(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 April 2002 (11.04.2002)

PCT

(10) International Publication Number
WO 02/29032 A2

(51) International Patent Classification⁷: C12N 15/00

(21) International Application Number: PCT/US01/31004

(22) International Filing Date: 1 October 2001 (01.10.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

09/677,584	30 September 2000 (30.09.2000)	US
60/279,702	28 March 2001 (28.03.2001)	US
PCT/US01/19367	14 June 2001 (14.06.2001)	US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier applications:

US	US01/19367 (CIP)
Filed on	14 June 2001 (14.06.2001)
US	60/279,702 (CIP)
Filed on	28 March 2001 (28.03.2001)
US	09/677,584 (CIP)
Filed on	30 September 2000 (30.09.2000)

(71) Applicant *(for all designated States except US)*: **DIVERSA CORPORATION** [US/US]; 4955 Directors Place, San Diego, CA 92121 (US).

(72) **Inventors; and**

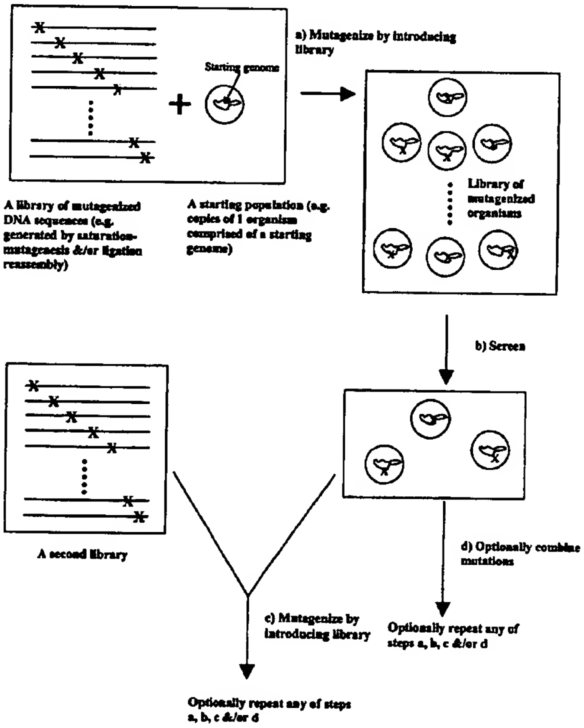
(75) **Inventors/Applicants** *(for US only)*: **SHORT, Jay, M.** [US/US]; P.O. Box 7214, Rancho Santa Fe, CA 92067-7214 (US). **FU, Pengcheng** [AU/US]; 7588 Charmant Drive #1914, San Diego, CA 92122-5079 (US). **LATTERICH, Martin** [DE/US]; 12539 Motellano Terrace, San Diego, CA 92130 (US). **WEI, Jing** [CN/US]; 10725 Wexford St. #6, San Diego, CA 92131 (US). **LEVIN, Michael** [RU/US]; 7565 Tupelo Cove, San Diego, CA 92126 (US).

(74) **Agents: EINHORN, Gregory, P.** et al.; Fish & Richardson P.C., 4350 La Jolla Village Drive, San Diego, CA 92122 (US).

(81) **Designated States** *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH,

[Continued on next page]

(54) **Title:** WHOLE CELL ENGINEERING BY MUTAGENIZING A SUBSTANTIAL PORTION OF A STARTING GENOME, COMBINING MUTATIONS, AND OPTIONALLY REPEATING


a) Mutagenize by introducing library
Starting genome
A library of mutagenized DNA sequences (e.g. generated by saturation-mutagenesis &/or ligation reassembly)
A starting population (e.g. copies of 1 organism comprised of a starting genome)
Library of mutagenized organisms
b) Screen
A second library
d) Optionally combine mutations
c) Mutagenize by introducing library
Optionally repeat any of steps a, b, c &/or d
Optionally repeat any of steps a, b, c &/or d

(57) **Abstract:** An invention comprising cellular transformation, directed evolution, and screening methods for creating novel transgenic organisms having desirable properties. In one embodiment, this invention provides a method of generating a transgenic organism, such as a microbe or a plant, having a plurality of traits that are differentially activatable. This invention also provides a method of retooling genes and gene pathways by the introduction of regulatory sequences, such as promoters, that are operable in an intended host, this conferring operability to a novel gene pathway when it is introduced into an intended host. For example a novel man-made gene pathway, generated based on microbially-derived progenitor templates, that is operable in a plant cell. This invention also provides a method of generating novel host organisms having increased expression of desirable traits, recombinant genes, and gene products. This invention provides novel methods for determining polypeptide profiles, and protein expression variations, which methods are applicable to all sample types disclosed herein. The present invention provides methods of simultaneously identifying and quantifying individual proteins in complex protein mixtures. Additionally this invention provides methods for cellular and metabolic engineering of new and modified phenotypes by using "on-line" or "real-time" metabolic flux analysis.

GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) **Designated States** *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WHOLE CELL ENGINEERING BY MUTAGENIZING A SUBSTANTIAL PORTION OF A STARTING GENOME, COMBINING MUTATIONS, AND OPTIONALLY REPEATING

A - FIELD OF THE INVENTION

This invention relates to the field of cellular and whole organism engineering. Specifically, this invention relates to a cellular transformation, directed evolution, and screening method for creating novel transgenic organisms having desirable properties. Thus in one aspect, this invention relates to a method of generating a transgenic organism, such as a microbe or a plant, having a plurality of traits that are differentially activatable.

This invention also relates to the field of protein engineering. Specifically, this invention relates to a directed evolution method for preparing a polynucleotide encoding a polypeptide. More specifically, this invention relates to a method of using mutagenesis to generate a novel polynucleotide encoding a novel polypeptide, which novel polypeptide is itself an improved biological molecule &/or contributes to the generation of another improved biological molecule. More specifically still, this invention relates to a method of performing both non-stochastic polynucleotide chimerization and non-stochastic site-directed point mutagenesis.

Thus, in one aspect, this invention relates to a method of generating a progeny set of chimeric polynucleotide(s) by means that are synthetic and non-stochastic, and where the design of the progeny polynucleotide(s) is derived by analysis of a parental set of polynucleotides &/or of the polypeptides correspondingly encoded by the parental polynucleotides. In another aspect this invention relates to a method of performing site-directed mutagenesis using means that are exhaustive, systematic, and non-stochastic.

Furthermore this invention relates to a step of selecting from among a generated set of progeny molecules a subset comprised of particularly desirable species, including by a process termed end-selection, which subset may then be screened further. This invention also relates to the step of screening a set of polynucleotides for the production of a polypeptide &/or of another expressed biological molecule having a useful property.

Novel biological molecules whose manufacture is taught by this invention include genes, gene pathways, and any molecules whose expression is affected thereby, including directly encoded polypetides &/or any molecules affected by such polypeptides. Said novel biological molecules include those that contain a carbohydrate, a lipid, a nucleic acid, &/or a protein component, and specific but non-limiting examples of these include antibiotics, antibodies, enzymes, and steroidal and non-steroidal hormones.

In a particular non-limiting aspect, the present invention relates to enzymes, particularly to thermostable enzymes, and to their generation by directed evolution. More particularly, the present invention relates to thermostable enzymes which are stable at high temperatures and which have improved activity at lower temperatures.

B - BACKGROUND

General Overview of the Problem to Be Solved

Brief Summary: It is instantly appreciated that the process of performing a genetic manipulation on a organism to achieve a genetic alteration, whether it is on a unicellular or on a multi-cellular organism, can lead to harmful, toxic, noxious, or even lethal effects on the manipulated organism. This is particularly true when the genetic manipulation becomes sizable. From a technical point of view, this problem is seen as one of the current obstacles that hinder the creation of genetically altered organisms having a large number of transgenic traits.

On the marketing side, is instantly appreciated that the purchase price of a genetically altered organism is often dictated by, or proportional to, the number of transgenic traits that have been introduced into the organism. Consequently, a genetically altered organism having a large number of stacked transgenic traits can be quite costly to produce and purchase and economically in low demand.

On the other hand, the generation of organism having but a single genetically introduced trait can also lead to the incurrence of undesirable costs, although for other reasons. It is thus appreciated that the separate production, marketing, & storage of genetically altered organisms each having a single transgenic traits can incur costs, including inventory costs, that are undesirable. For example, the storage of such organisms may require a separate bin to be used for each trait. Furthermore, the value of an organisms having a single particular trait is often intimately tied to the marketability of that particular trait, and when that marketability diminishes, inventories of such organisms cannot be sold in other markets.

The instant invention solves these and other problems by providing a method of producing genetically altered organisms having a large number of stacked traits that are differentially activatable. Upon purchasing such a genetically altered organism (having a large number of differentially activatable stacked traits), the purchasing customer has the option of selecting and paying for particular traits among the total that can then be activated differentially. One economic advantage provided by this invention is that the storage of such genetically altered organisms is simplified since, for example, one bin could be used to store a large number of traits. Moreover, a single organism of this type can satisfy the demands for a variety of traits; consequently, such an organism can be sold in a variety of markets.

To achieve the production of genetically altered organisms having a large number of stacked traits that are differentially activatable, this invention provides - in one specific aspect – a process comprising the step of monitoring a cell or organism at holistic level. This serves as a way of collecting holistic - rather than isolated - information about a working cell or organism that is being subjected to a substantial amount of genetic manipulation. This invention further provides that this type of holistic monitoring can include the detection of all morphological, behavioral, and physical parameters.

Accordingly, the holistic monitoring provided by this invention can include the identification &/or quantification of all the genetic material contained in a working cell or organism (e.g. all nucleic acids including the entire genome, messenger RNA's, tRNA's, rRNA's, and mitochondrial nucleic acids, plasmids, phages, phagemids, viruses, as well as all episomal nucleic acids and endosymbiont nucleic acids). Furthermore this invention provides that this type of holistic monitoring can include all gene products produced by the working cell or organisms.

Furthermore, the holistic monitoring provided by this invention can include the identification &/or quantification of all molecules that are chemically at least in part protein in a working cell or organism. The holistic monitoring provided by this invention can also include the identification &/or quantification of all molecules that are chemically at least in part carbohydrate in a working cell or organism. The holistic monitoring provided by this invention can also include the identification &/or quantification of all molecules that are chemically at least in part proteoglycan in a working cell or organism. The holistic monitoring provided by this invention can also include the identification &/or quantification of all molecules that are chemically at least in part glycoprotein in a working cell or organism. The holistic monitoring provided by this invention can also include the identification &/or quantification of all molecules that are chemically at least in part nucleic acids in a working cell or organism. The holistic monitoring provided by this invention can also include the identification &/or quantification of all molecules that are chemically at least in part lipids in a working cell or organism.

In one aspect, this invention provides that the ability to differentially activate a trait from among many, such as a enzyme from among many enzymes, depends the enzyme(s) to be activated having a unique activity profile (or activity fingerprint). An enzyme's activity profile includes the reaction(s) it catalyzes and its specificity. Thus, an enzymes activity profile includes its:

- Catalyzed reaction(s)
- Reaction type
- Natural substrate(s)
- Substrate spectrum
- Product spectrum
- Inhibitor(s)
- Cofactor(s)/prostetic group(s)
- Metal compounds/salts that affect it
- Turnover number
- Specific activity
- Km value
- pH optimum
- pH range
- Temperature optimum
- Temperature range

It is also instantly appreciated that enzymes are differentially affected by exposure to varying degrees of processing (e.g. upon extraction &/or purification) and exposure (e.g. to suboptimal storage conditions). Accordingly, enzyme differences may surface after exposure to:

- Isolation/Preparation
- Purification
- Crystallization
- Renaturation

It is instantly appreciated that differences in molecular stability can also be used advantageously to differentially activate or inactivate selected enzymes, by exposing the enzymes for an appropriate time to variations in:

- pH
- Temperature
- Oxidation
- Organic solvent(s)
- Miscellaneous storage conditions

It is thus appreciated that in order to be able to differentially activate selected traits among a plurality of stacked traits, it is desirable to introduce into a working cell or organism traits conferred by molecules (e.g. enzymes) having very unique profiles (e.g. unique enzyme fingerprints). Furthermore , it is appreciated that in order to obtain the molecules having a representation of a wide range of molecular fingerprints, it is advantageous to harvest molecules from the widest possible reaches nature's diversity. Thus, it is beneficial to harvest molecules not only from cultured mesophilic organisms, but also from extremophiles that are largely uncultured.

In another aspect, it is instantly appreciated that harvesting the full potential of nature's diversity can include both the step of discovery and the step of optimizing what is discovered. For example, the step of discovery allows one to mine biological molecules that have commercial utility. It is instantly appreciated that the ability to harvest the full richness of biodiversity, i.e. to mine biological molecules from a wide range of environmental conditions, is critical to the ability to discover novel molecules adapted to funtion under a wide variety of conditions, including extremes of conditions, such as may be found in a commercial application.

However, it is also instantly appreciated that only occassionally are there criteria for selection &/or survival in nature that point in the exact direction of particular commercial needs. Instead, it is often the case that a naturally occurring molecule will require a certain amount of change – from fine tuning to sweeping modification – in order to fulfill a particular unmet commercial need. Thus, to meet certain commercial needs (e.g., a need for a molecule that is fucntional under a specific set of commercial processing conditions) it is sometimes advantageous to experimentally modify a naturally expresed molecule to achieve properties beyond what natural evolution has provided &/or is likely to provide in the near future.

The approach, termed directed evolution, of experimentally modifying a biological molecule towards a desirable property, can be achieved by mutagenizing one or more parental molecular templates and by idendifying any desirable molecules among the progeny molecules. Currently available technologies in directed evolution include methods for achieving stochastic (i.e. random) mutagenesis and methods for achieving non-stochastic (non-random) mutagenesis. However, critical shortfalls in both types of methods are identified in the instant disclosure.

In prelude, it is noteworthy that it may be argued philosophically by some that all mutagenesis – if considered from an objective point of view – is non-stochastic; and furthermore that the entire universe is undergoing a process that – if considered from an objective point of view – is non-stochastic. Whether this is true is outside of the scope of the instant consideration. Accordingly, as used herein, the terms "randomness", "uncertainty", and "unpredictability" have subjective meanings, and the knowledge, particularly the predictive knowledge, of the designer of an experimental process is a determinant of whether the process is stochastic or non-stochastic.

By way of illustration, stochastic or random mutagenesis is exemplified by a situation in which a progenitor molecular template is mutated (modified or changed) to yield a set of progeny molecules having mutation(s) that are not predetermined. Thus, in an in vitro stochastic mutagenesis reaction, for example, there is not a particular predetermined product whose production is intended; rather there is an uncertainty – hence randomness – regarding the exact nature of the mutations achieved, and thus also regarding the products generated. In contrast, non-stochastic or non-random mutagenesis is exemplified by a situation in which a progenitor molecular template is mutated (modified or changed) to yield a progeny molecule having one or more predetermined mutations. It is appreciated that the presence of background products in some quantity is a reality in many reactions where molecular processing occurs, and the presence of these background products does not detract from the non-stochastic nature of a mutagenesis process having a predetermined product.

Thus, as used herein, stochastic mutagenesis is manifested in processes such as error-prone PCR and stochastic shuffling, where the mutation(s) achieved are random or not predetermined. In contrast, as used herein, non-stochastic mutagenesis is manifested in instantly disclosed processes such as gene site-saturation mutagenesis and synthetic ligation reassembly, where the exact chemical structure(s) of the intended product(s) are predetermined.

In brief, existing mutagenesis methods that are non-stochastic have been serviceable in generating from one to only a very small number of predetermined mutations per method application, and thus produce per method application from one to only a few progeny molecules that have predetermined molecular structures. Moreover, the types of mutations currently available by the application of these non-stochastic methods are also limited, and thus so are the types of progeny mutant molecules.

In contrast, existing methods for mutagenesis that are stochastic in nature have been serviceable for generating somewhat larger numbers of mutations per method application – though in a random fashion & usually with a large but unavoidable contingency of undesirable background products. Thus, these existing stochastic methods can produce per method application larger numbers of progeny molecules, but that have undetermined molecular structures. The types of mutations that can be achieved by application of these current stochastic methods are also limited, and thus so are the types of progeny mutant molecules.

It is instantly appreciated that there is a need for the development of non-stochastic mutagenesis methods that:

1) Can be used to generate large numbers of progeny molecules that have predetermined molecular structures;

2) Can be used to readily generate more types of mutations;

3) Can produce a correspondingly larger variety of progeny mutant molecules;

4) Produce decreased unwanted background products;

5) Can be used in a manner that is exhaustive of all possibilities; and

6) Can produce progeny molecules in a systematic & non-repetitive way.

The instant invention satisfies all of these needs.

Directed Evolution Supplements Natural Evolution: Natural evolution has been a springboard for directed or experimental evolution, serving both as a reservoir of methods to be mimicked and of molecular templates to be mutagenized. It is appreciated that, despite its intrinsic process-related limitations (in the types of favored &/or allowed mutagenesis processes) and in its speed, natural evolution has had the advantage of having been in process for millions of years & and throughout a wide diversity of environments. Accordingly, natural evolution (molecular mutagenesis and selection in nature) has resulted in the generation of a wealth of biological compounds that have shown usefulness in certain commercial applications.

However, it is instantly appreciated that many unmet commercial needs are discordant with any evolutionary pressure &/or direction that can be found in nature. Moreover, it is often the case that when commercially useful mutations would otherwise be favored at the molecular level in nature, natural evolution often overrides the positive selection of such mutations, e.g. when there is a concurrent detriment to an organism as a whole (such as when a favorable mutation is accompanied by a detrimental mutation).

Additionally, natural evolution is often slow, and favors fidelity in many types of replication. Additionally still, natural evolution often favors a path paved mainly by consecutive beneficial mutations while tending to avoid a plurality of successive negative mutations, even though such negative mutations may prove beneficial when combined, or may lead - through a circuitous route - to final state that is beneficial.

Moreover, natural evolution advances through specific steps (e.g. specific mutagenesis and selection processes), with avoidance of less favored steps. For example, many nucleic acids do not reach close enough proximity to each other in a operative environment to undergo chimerization or incorporation or other types of transfers from one species to another. Thus, e.g., when sexual intercourse between 2 particular species is avoided in nature, the chimerization of nucleic acids from these 2 species is likewise unlikely, with parasites common to the two species serving as an example of a very slow passageway for inter-molecular encounters and exchanges of DNA. For another example, the generation of a molecule causing self-toxicity or self-lethality or sexual sterility is avoided in nature. For yet another example, the propagation of a molecule having no particular immediate benefit to an organism is prone to vanish in subsequent generations of the organism. Furthermore, e.g., there is no selection pressure for improving the performance of molecule under conditions other than those to which it is exposed in its endogenous environment; e.g. a cytoplasmic molecule is not likely to acquire functional features extending beyond what is required of it in the cytoplasm. Furthermore still, the propagation of a biological molecule is susceptible to any global detrimental effects – whether caused by itself or not – on its ecosystem. These and other characteristics greatly limit the types of mutations that can be propagated in nature.

On the other hand, directed (or experimental) evolution – particularly as provided herein – can be performed much more rapidly and can be directed in a more streamlined manner at evolving a predetermined molecular property that is commercially desirable where nature does not provide one &/or is not likely to provide. Moreover, the directed evolution invention provided herein can provide more wide-ranging possibilities in the types of steps that can be used in mutagenesis and selection processes. Accordingly, using templates harvested from nature, the instant directed evolution invention provides more wide-ranging possibilities in the types of progeny molecules that can be generated and in the speed at which they can be generated than often nature itself might be expected to in the same length of time.

In a particular exemplification, the instantly disclosed directed evolution methods can be applied iteratively to produce a lineage of progeny molecules (e.g. comprising successive sets of progeny molecules) that would not likely be propagated (i.e., generated &/or selected for) in nature, but that could lead to the generation of a desirable downstream mutagenesis product that is not achievable by natural evolution.

Previous Directed Evolution Methods Are Suboptimal:

Mutagenesis has been attempted in the past on many occasions, but by methods that are inadequate for the purpose of this invention. For example, previously described non-stochastic methods have been serviceable in the generation of only very small sets of progeny molecules (comprised often of merely a solitary progeny molecule). By way of illustration, a chimeric gene has been made by joining 2 polynucleotide fragments using compatible sticky ends generated by restriction enzyme(s), where each fragment is derived from a separate progenitor (or parental) molecule. Another example might be the mutagenesis of a single codon position (i.e. to achieve a codon substitution, addition, or deletion) in a parental polynucleotide to generate a single progeny polynucleotide encoding for a single site-mutagenized polypeptide.

Previous non-stochastic approaches have only been serviceable in the generation of but one to a few mutations per method application. Thus, these previously described non-stochastic methods thus fail to address one of the central goals of this invention, namely the exhaustive and non-stochastic chimerization of nucleic acids. Accordingly previous non-stochastic methods leave untapped the vast majority of the possible point mutations, chimerizations, and combinations thereof, which may lead to the generation of highly desirable progeny molecules.

In contrast, stochastic methods have been used to achieve larger numbers of point mutations and/or chimerizations than non-stochastic methods; for this reason, stochastic methods have comprised the predominant approach for generating a set of progeny molecules that can be subjected to screening, and amongst which a desirable molecular species might hopefully be found. However, a major drawback of these approaches is that – because of their stochastic nature – there is a randomness to the exact components in each set of progeny molecules that is produced. Accordingly, the experimentalist typically has little or no idea what exact progeny molecular species are represented in a particular reaction vessel prior to their generation. Thus, when a stochastic procedure is repeated (e.g. in a continuation of a search for a desirable progeny molecule), the re-generation and re-screening of previously discarded undesirable molecular species becomes a labor-intensive obstruction to progress, causing a circuitous – if not circular – path to be taken. The drawbacks of such a highly suboptimal path can be addressed by subjecting a stochastically generated set of progeny molecules to a labor-incurring process, such as sequencing, in order to identify their molecular structures, but even this is an incomplete remedy.

Moreover, current stochastic approaches are highly unsuitable for comprehensively or exhaustively generating all the molecular species within a particular grouping of mutations, for attributing functionality to specific structural groups in a template molecule (e.g. a specific single amino acid position or a sequence comprised of two or more amino acids positions), and for categorizing and comparing specific grouping of mutations. Accordingly, current stochastic approaches do not inherently enable the systematic elimination of unwanted mutagenesis results, and are, in sum, burdened by too many inherently shortcomings to be optimal for directed evolution.

In a non-limiting aspect, the instant invention addresses these problems by providing non-stochastic means for comprehensively and exhaustively generating all possible point mutations in a parental template. In another non-limiting aspect, the instant invention further provides means for exhaustively generating all possible chimerizations within a group of chimerizations. Thus, the aforementioned problems are solved by the instant invention.

Specific shortfalls in the technological landscape addressed by this invention include:

1) Site-directed mutagenesis technologies, such as sloppy or low-fidelity PCR, are ineffective for systematically achieving at each position (site) along a polypeptide sequence the full (saturated) range of possible mutations (i.e. all possible amino acid substitutions).

2) There is no relatively easy systematic means for rapidly analyzing the large amount of information that can be contained in a molecular sequence and in the potentially colossal number or progeny molecules that could be conceivably obtained by the directed evolution of one or more molecular templates.

3) There is no relatively easy systematic means for providing comprehensive empirical information relating structure to function for molecular positions.

4) There is no easy systematic means for incorporating internal controls, such as positive controls, for key steps in certain mutagenesis (e.g. chimerization) procedures.

5) There is no easy systematic means to select for a specific group of progeny molecules, such as full-length chimeras, from among smaller partial sequences.

An exceedingly large number of possibilities exist for the purposeful and random combination of amino acids within a protein to produce useful hybrid proteins and their corresponding biological molecules encoding for these hybrid proteins, i.e., DNA, RNA. Accordingly, there is a need to produce and screen a wide variety of such hybrid proteins for a desirable utility, particularly widely varying random proteins.

The complexity of an active sequence of a biological macromolecule (e.g., polynucleotides, polypeptides, and molecules that are comprised of both polynucleotide and polypeptide sequences) has been called its information content ("IC"), which has been defined as the resistance of the active protein to amino acid sequence variation (calculated from the minimum number of invariable amino acids (bits) required to describe a family of related sequences with the same function). Proteins that are more sensitive to random mutagenesis have a high information content.

Molecular biology developments, such as molecular libraries, have allowed the identification of quite a large number of variable bases, and even provide ways to select functional sequences from random libraries. In such libraries, most residues can be varied (although typically not all at the same time) depending on compensating changes in the context. Thus, while a 100 amino acid protein can contain only 2,000 different mutations, 20^{100} sequence combinations are possible.

Information density is the IC per unit length of a sequence. Active sites of enzymes tend to have a high information density. By contrast, flexible linkers of information in enzymes have a low information density.

Current methods in widespread use for creating alternative proteins in a library format are error-prone polymerase chain reactions and cassette mutagenesis, in which the specific region to be optimized is replaced with a synthetically mutagenized oligonucleotide. In both cases, a substantial number of mutant sites are generated around certain sites in the original sequence.

Error-prone PCR uses low-fidelity polymerization conditions to introduce a low level of point mutations randomly over a long sequence. In a mixture of fragments of unknown sequence, error-prone PCR can be used to mutagenize the mixture. The published error-prone PCR protocols suffer from a low processivity of the polymerase. Therefore, the protocol is unable to result in the random mutagenesis of an average-sized gene. This inability limits the practical application of error-prone PCR. Some computer simulations have suggested that point mutagenesis alone may often be too gradual to allow the large-scale block changes that are required for continued and dramatic sequence evolution. Further, the published error-prone PCR protocols do not allow for amplification of DNA fragments greater than 0.5 to 1.0 kb, limiting their practical application. In addition, repeated cycles of error-prone PCR can lead to an accumulation of neutral mutations with undesired results, such as affecting a protein's immunogenicity but not its binding affinity.

In oligonucleotide-directed mutagenesis, a short sequence is replaced with a synthetically mutagenized oligonucleotide. This approach does not generate combinations of distant mutations and is thus not combinatorial. The limited library size relative to the vast sequence length means that many rounds of selection are unavoidable for protein optimization. Mutagenesis with synthetic oligonucleotides requires sequencing of individual clones after each selection round followed by grouping them into families, arbitrarily choosing a single family, and reducing it to a consensus motif. Such motif is re-synthesized and reinserted into a single gene followed by additional selection. This step process constitutes a statistical bottleneck, is labor intensive, and is not practical for many rounds of mutagenesis.

Error-prone PCR and oligonucleotide-directed mutagenesis are thus useful for single cycles of sequence fine-tuning, but rapidly become too limiting when they are applied for multiple cycles.

Another limitation of error-prone PCR is that the rate of down-mutations grows with the information content of the sequence. As the information content, library size, and mutagenesis rate increase, the balance of down-mutations to up-mutations will statistically prevent the selection of further improvements (statistical ceiling).

In cassette mutagenesis, a sequence block of a single template is typically replaced by a (partially) randomized sequence. Therefore, the maximum information content that can be obtained is statistically limited by the number of random sequences (i.e., library size). This eliminates other sequence families which are not currently best, but which may have greater long term potential.

Also, mutagenesis with synthetic oligonucleotides requires sequencing of individual clones after each selection round. Thus, such an approach is tedious and impractical for many rounds of mutagenesis.

Thus, error-prone PCR and cassette mutagenesis are best suited, and have been widely used, for fine-tuning areas of comparatively low information content. One apparent exception is the selection of an RNA ligase ribozyme from a random library using many rounds of amplification by error-prone PCR and selection.

In nature, the evolution of most organisms occurs by natural selection and sexual reproduction. Sexual reproduction ensures mixing and combining of the genes in the offspring of the selected individuals. During meiosis, homologous chromosomes from the parents line up with one another and cross-over part way along their length, thus randomly swapping genetic material. Such swapping or shuffling of the DNA allows organisms to evolve more rapidly.

In recombination, because the inserted sequences were of proven utility in a homologous environment, the inserted sequences are likely to still have substantial information content once they are inserted into the new sequence.

Theoretically there are 2,000 different single mutants of a 100 amino acid protein. However, a protein of 100 amino acids has 20^{100} possible sequence combinations, a number which is too large to exhaustively explore by conventional methods. It would be advantageous to develop a system which would allow generation and screening of all of these possible combination mutations.

Some workers in the art have utilized an *in vivo* site specific recombination system to generate hybrids of combine light chain antibody genes with heavy chain antibody genes for expression in a phage system. However, their system relies on specific sites of recombination and is limited accordingly. Simultaneous mutagenesis of antibody CDR regions in single chain antibodies (scFv) by overlapping extension and PCR have been reported.

Others have described a method for generating a large population of multiple hybrids using random *in vivo* recombination. This method requires the recombination of two different libraries of plasmids, each library having a different selectable marker. The method is limited to a finite number of recombinations equal to the number of selectable markers existing, and produces a concomitant linear increase in the number of marker genes linked to the selected sequence(s).

In vivo recombination between two homologous, but truncated, insect-toxin genes on a plasmid has been reported as a method of producing a hybrid gene. The *in vivo* recombination of substantially mismatched DNA sequences in a host cell having defective mismatch repair enzymes, resulting in hybrid molecule formation has been reported.

C - SUMMARY OF THE INVENTION

This invention relates generally to the field of cellular and whole organism engineering. Specifically, this invention relates to a cellular transformation, directed evolution, and screening method for creating novel transgenic organisms having desirable properties. Thus in one aspect, this invention relates to a method of generating a transgenic organism, such as a microbe or a plant, having a plurality of traits that are differentially activatable.

In one embodiment, this invention is directed to a method of producing an improved organism having a desirable trait to by: a) obtaining an initial population of organisms, b) generating a set of mutagenized organisms, such that when all the genetic mutations in the set of mutagenized organisms are taken as a whole, there is represented a set of substantial genetic mutations, and c) detecting the presence of said improved organism. This invention provides that any of steps a), b), and c) can be further repeated in any particular order and any number of times; accordingly, this invention specifically provides methods comprised of any iterative combination of steps a), b), and c), with a number of iterations.

In another embodiment, this invention is directed to a method of producing an improved organism having a desirable trait to by: a) obtaining an initial population of organisms, which can be a clonal population or otherwise, b) generating a set of mutagenized organisms each having at least one genetic mutation, such that when all the genetic mutations in the set of mutagenized organisms are taken as a whole, there is represented a set of substantial genetic mutations c) detecting the manifestation of at least two genetic mutations, and d) introducing at least two detected genetic mutations into one organism. Additionally, this invention provides that any of steps a), b), c), and d) can be further repeated in any particular order and any number of times; accordingly, this invention specifically provides methods comprised of any iterative combination of steps a), b), c), and d), with a total number of iterations can be from one up to one million, including specifically every integer value in between.

In a preferred aspect of embodiments specified herein the step of b) generating a second set of mutagenized organisms is comprised of generating a plurality of organisms, each of which organisms has a particular transgenic mutation.

As used herein, "**generating a set of mutagenized organisms having genetic mutations**" can be achieved by any means known in the art to mutagenized including any radiation known to mutagenized, such as ionizing and ultra violet. Further examples of serviceable mutagenizing methods include site-saturation mutagenesis, transposon-based methods, and homologous recombination.

"**Combining**" means incorporating a plurality of different genetic mutations in the genetic makeup (e.g. the genome) of the same organism; and methods to achieve this "combining" step including sexual recombination, homologous recombination, and transposon-based methods.

As used herein, an "**initial population of organisms**" means a "**working population of organisms**", which refers simply to a population of organisms with which one is working, and which is comprised of at least one organism. An "**initial population of organisms**" which can be a clonal population or otherwise.

Accordingly, in step 1) an "**initial population of organisms**" may be a population of multicellular organisms or of unicellular organisms or of both. An "initial population of organisms" may be comprised of unicellular organisms or multicellular organisms or both. An "initial population of organisms" may be comprised of prokaryotic organisms or eukaryotic organisms or both. This invention provides that an "initial population of organisms" is comprised of at least one organism, and preferred embodiments include at least that .

By "**organism**" is meant any biological form or thing that is capable of self replication or replication in a host. Examples of "organisms" include the following kinds of organisms (which kinds are not necessarily mutually-exclusive): animals, plants, insects, cyanobacteria, microorganisms, fungi, bacteria, eukaryotes, prokaryotes, mycoplasma, viral organisms (including DNA viruses, RNA viruses), and prions.

Non-limiting particularly preferred examples of kinds of "**organisms**" also include Archaea (archaebacteria) and Bacteria (eubacteria). Non-limiting examples of Archaea (archaebacteria) include Crenarchaeota, Euryarchaeota, and Korarchaeota. Non-limiting examples Bacteria (eubacteria) include Aquificales, CFB/Green sulfur bacteria group, Chlamydiales/Verrucomicrobia group, Chrysiogenes group, Coprothermobacter group, Cyanobacteria & chloroplasts, Cytophaga/Flexibacter /Bacteriods group, Dictyoglomus group, Fibrobacter/Acidobacteria group, Firmicutes, Flexistipes group, Fusobacteria, Green non-sulfur bacteria, Nitrospira group, Planctomycetales, Proteobacteria, Spirochaetales, Synergistes group, Thermodesulfobacterium group, Thermotogales, Thermus/Deinococcus group. As non-limiting examples, particularly preferred kinds of organisms include Aquifex, Aspergillus, Bacillus, Clostridium, E. coli, Lactobacillus, Mycobacterium, Pseudomonas, Streptomyces, and Thermotoga. As additional non-limiting examples, particularly preferred organisms include cultivated organisms such as CHO, VERO, BHK, HeLa, COS, MDCK, Jurkat, HEK-293, and WI38. Particularly preferred non-limiting examples of organisms further include host organisms that are serviceable for the expression of recombinant molecules. Organisms further include primary cultures (e.g. cells from harvested mammalian tissues), immortalized cells, all cultivated and culturable cells and multicellular organisms, and all uncultivated and uculturable cells and multicellular organisms.

In a preferred embodiment, knowledge of genomic information is useful for performing the claimed methods; thus, this invention provides the following as preferred but non-limiting examples of organisms that are particularly serviceable for this invention, because there is a significant amount of - if not complete - genomic sequence information (in terms of primary sequence &/or annotation) for these organisms: Human, Insect (e.g. *Drosophila melanogaster*), Higher plants (e.g. *Arabidopsis thaliana*), Protozoan (e.g. *Plasmodium falciparum*), Nematode (e.g. *Caenorhabditis elegans*), Fungi(e.g. *Saccharomyces cerevisiae*), Proteobacteria gamma subdivision (e.g. *Escherichia coli* K-12 , *Haemophilus influenzae* Rd, *Xylella fastidiosa* 9a5c, *Vibrio cholerae* El Tor N16961, *Pseudomonas aeruginosa* PA01, *Buchnera* sp. APS), Proteobacteria beta subdivision (e.g. *Neisseria meningitidis* MC58 (serogroup B), *Neisseria meningitidis* Z2491 (serogroup A)), Proteobacteria other subdivisions (e.g. *Helicobacter pylori* 26695, *Helicobacter pylori* J99, *Campylobacter jejuni* NCTC11168, *Rickettsia prowazekii*), Gram-positive bacteria (e.g. *Bacillus subtilis, Mycoplasma genitalium, Mycoplasma pneumoniae, Ureaplasma urealyticum, Mycobacterium tuberculosis* H37Rv), Chlamydia (e.g. *Chlamydia trachomatisserovar* D, *Chlamydia muridarum* (*Chlamydia trachomatis* MoPn), *Chlamydia pneumoniae* CWL029, *Chlamydia pneumoniae* AR39, *Chlamydia pneumoniae* J138), Spirochete (e.g. *Borrelia burgdorferi* B31, *Treponema pallidum*), Cyanobacteria (e.g. *Synechocystis* sp. PCC6803), Radioresistant bacteria (e.g. *Deinococcus radiodurans* R1), Hyperthermophilic bacteria (e.g. *Aquifex aeolicus* VF5, *Thermotoga maritima* MSB8), and Archaea (e.g. *Methanococcus jannaschii, Methanobacterium thermoautotrophicum* deltaH, *Archaeoglobus fulgidus, Pyrococcus horikoshii* OT3, *Pyrococcus abyssi, Aeropyrum pernix* K1).

Non-limiting particularly preferred examples of kinds of plant "organisms" include those listed in Table 1.

Table 1. Non-limiting examples of plant organisms and sources of transgenic molecules (e.g. nucleic acids & nucleic acid products)

1. Alfalfa	39. Pepper
2. Amelanchier laevis	40. Persimmon
3. Apple	41. Petunia
4. Arab. thaliana	42. Pine
5. Arabidopsis	43. Pineapple
6. Aspergillus flavus	44. Pink bollworm
7. Barley	45. Plum
8. Beet	46. Poplar
9. Belladonna	47. Potato
10. Brassica oleracea	48. Pseudomonas
11. Carrot	49. Pseudomonas putida
12. Chrysanthemum	50. Pseudomonas syringae
13. Cichorium intybus	51. Rapeseed
14. Clavibacter	52. Rhizobium
15. Clavibacter xyli	53. Rhizobium etli
16. Coffee	54. Rhizobium fredii
17. Corn	55. Rhizobium leguminosarum
18. Cotton	56. Rhizobium meliloti
19. Cranberry	57. Rice
20. Creeping bentgrass	58. Rubus idaeus
21. Cryphonectria parasitica	59. Spruce
22. Eggplant	60. Soybean
23. Festuca arundinacea	61. Squash
24. Fusarium graminearum	62. Squash-cucumber

25. Fusarium moniliforme	63. Squash-cucurbita texana
26. Fusarium sporotrichioides	64. Strawberry
27. Gladiolus	65. Sugarcane
28. Grape	66. Sunflower
29. Heterorhabditis bacteriophora	67. Sweet potato
30. Kentucky bluegrass	68. Sweetgum
31. Lettuce	69. TMV
32. Melon	70. Tobacco
33. Oat	71. Tomato
34. Onion	72. Walnut
35. Papaya	73. Watermelon
36. Pea	74. Wheat
37. Peanut	75. Xanthomonas
38. Pelargonium	76. Xanthomonas campestris

As used herein, the meaning of "**generating a set of mutagenized organisms having genetic mutations**" includes the steps of substituting, deleting, as well as introducing a nucleotide sequence into organism; and this invention provides a nucleotide sequence that serviceable for this purpose may be a single-stranded or double-stranded and the fact that its length may be from one nucleotide up to 10,000,000,000 nucleotides in length including specifically every integer value in between.

A mutation in an organism includes any alteration in the structure of one or more molecules that encode the organism. These molecules include nucleic acid, DNA, RNA, prionic molecules, and may be exemplified by a variety of molecules in an organism such as a DNA that is genomic, episomal, or nucleic, or by a nucleic acid that is vectoral (e.g. viral, cosmid, phage, phagemid).

In one aspect, as used herein, a "**set of substantial genetic mutations**" is preferably a disruption (e.g. a functional knock-out) of at least about 15 to about 150,000 genomic locations or nucleotide sequences (e.g. genes, promoters, regulatory sequences, codons etc.), including specifically every integer value in between. In another aspect, as used herein, a "**set of substantial genetic mutations**" is preferably an alteration in an expression level (e.g. decreased or increased expression level) or an alteration in the expression pattern (e.g. throughout a period of time) of at least about 15 to about 150,000 genes, including specifically every integer value in between. Corresponding to another aspect, as used herein, a "**set of substantial genetic mutations**" is preferably an alteration in an expression level (e.g. decreased or increased expression level) or an alteration in the expression pattern (e.g. throughout a period of time) of at least about 15 to about 150,000 gene products &/or phenotypes &/or traits, including specifically every integer value in between.

In another aspect, as used herein, a "**set of substantial genetic mutations**" with respect to an organism (or type of organism) is preferably a disruption (e.g. a functional knock-out) of at least about 1% to about 100% of genomic locations or nucleotide sequences (e.g. genes, promoters, regulatory sequences, codons etc.) in the organism (or type of organism), including specifically percentages of every integer value in between. In another aspect, as used herein, a "**set of substantial genetic mutations**" is preferably an alteration in an expression level (e.g. decreased or increased expression level) or an alteration in the expression pattern (e.g. throughout a period of time) of at least about 1% to about 100% of genes in an organism (or type of organism), including specifically percentages of every integer value in between. Corresponding to another aspect, as used herein, a "**set of substantial genetic mutations**" is preferably an alteration in an expression level (e.g. decreased or increased expression level) or an alteration in the expression pattern (e.g. throughout a period of time) of at least about 1% to about 100% of the gene products &/or phenotypes &/or traits of an organism (or type of organism), including specifically every integer value in between.

In yet another aspect, as used herein, a "**set of substantial genetic mutations**" is preferably an introduction or deletion of at least about 15 to 150,000 genes promoters or other nucleotide sequences (where each sequence is from 1 base to 10,000,000 bases), including specifically every integer value in between. For example, one can introduce a library of at least about 15 to 150,000 nucleotides (genes or promoters) produced by "site-saturation mutagenesis" &/or by "ligation reassembly" (including any specific aspect thereof provided herein) into an "**initial population of organisms**".

It is provided that wherever the manipulation of a plurality of "**genes**" is mentioned herein, gene pathways (e.g. that ultimately lead to the production of small molecules) are also included. It is appreciated herein that knocking-out, altering expression level, and altering expression pattern can be achieved, by non-limiting exemplification, by mutagenizing a nucleotide sequence corresponding gene as well as a corresponding promoter that affects the expression of the gene.

As used herein, a "**mutagenized organism**" includes any organism that has been altered by a genetic mutation.

A "**genetic mutation**" can be, by way of non-limiting and non-mutually exclusive exemplification, and change in the nucleotide sequence (DNA or RNA) with respect to genomic, extra-genomic, episomal, mitochondrial, and any nucleotide sequence associated with (e.g. contained within or considered part of) an organism..

According to this invention, detecting the manifestation of a "**genetic mutation**" means "**detecting the manifestation of a detectable parameter**", including but not limited to a change in the genomic sequence. Accordingly, this invention provides that a step of sequencing (&/or annotating) of and organism's genomic DNA is necessary for some methods of this invention, and exemplary but non-limiting aspects of this sequencing (&/or annotating) step are provided herein.

A detectable "**trait**", as used herein, is any detectable parameter associated with the organism. Accordingly, such a detectable "parameter" includes, by way of non-limiting exemplification, any detectable "nucleotide knock-in", any detectable "nucleotide knock-outs", any detectable "phenotype", and any detectable "genotype". By way of further illustration, a "**trait**" includes any substance produced or not produced by the organism. Accordingly, a "**trait**" includes viability or non-viability, behavior, growth rate, size, morphology. "**Trait**" includes increased (or alternatively decreased) expression of a gene product or gene pathway product. "**Trait**" also includes small molecule production (including vitamins, antibiotics), herbicide resistance, drought resistance, pest resistance, production of any recombinant biomolecule (ie.g. vaccines, enzymes, protein therapeutics, chiral enzymes). Additional examples of serviceable traits for this invention are shown in Table 2.

TABLE 2 – **Non-limiting examples of serviceable genes, gene products, phenotypes, or traits according to the methods of this invention (e.g. knockouts, knockins, increased or decreased expression level, increased or decreased expression pattern)**

Table 2 - Part 1. Non-limiting examples of genes or gene products

1.	17 kDa protein	53.	Cecropin
2.	3-hydroxy-3-methylglutaryl CoenzymeA reductase	54.	Cecropin B
3.	4-Coumarate:CoA ligase knockout	55.	Cellulose binding protein
4.	60 kDa protein	56.	Chalcone synthase knockout
5.	Ac transposable element	57.	Chitinase
6.	ACC deaminase	58.	Chitobiosidase
7.	ACC oxidase knockout	59.	Chloramphenicol acetyltransferase
8.	ACC synthase	60.	Cholera toxin B
9.	ACC synthase knockout	61.	Choline oxidase
10.	Acetohydroxyacid synthase variant	62.	Cinnamate 4-hydroxylase
11.	Acetolactate synthase	63.	Cinnamate 4-hydroxylase knockout
12.	Acetyl CoA carboxylase	64.	Coat protein
13.	ACP acyl-ACP thioesterase	65.	Coat protein knockout
14.	ACP thioesterase	66.	Conglycinin
15.	Acyl CoA reductase	67.	CryIA
16.	Acyl-ACP knockout	68.	CryIAb
17.	Acyl-ACP desaturase	69.	CryIAc
18.	Acyl-ACP desaturase knockout	70.	CryIB
19.	Acyl-ACP thioesterase	71.	CryIIA
20.	ADP glucose pyrophosphorylase	72.	CryIIIA
21.	ADP glucose pyrophosphorylase knockout	73.	CryVIA
22.	Agglutinin	74.	Cyclin dependent kinase
23.	Aleurone 1	75.	Cyclodextrin glycosyltransferase
24.	Alpha hordothinonin	76.	Cylindrical inclusion protein
25.	Alpha-amylase	77.	Cystathionine synthase
26.	Alpha-hemoglobin	78.	Delta-12 desaturase
27.	Aminoglycoside 3'-adenylytransferase	79.	Delta-12 desaturase knockout
28.	Amylase	80.	Delta-12 saturase
29.	Anionic peroxidase	81.	Delta-12 saturase knockout
30.	Antibody	82.	Delta-15 desaturase
31.	Antifungal protein	83.	Delta-15 desaturase knockout
32.	Antithrombin	84.	Delta-9 desaturase
33.	Antitrypsin	85.	Delta-9 desturase knockout
34.	Antiviral protein	86.	Deoxyhypusine synthase (DHS)
35.	Aspartokinase	87.	Deoxyhypusine synthase knockout
36.	Attacin E	88.	Diacylglycerol acetyl tansferase
37.	B1 regulatory gene	89.	Dihydrodipicolinate synthase
38.	B-1,3-glucanase knockout	90.	Dihydrofolate reductase
39.	B-1,4-endoglucanase knockout	91.	Diptheria toxin A
40.	Bacteropsin	92.	Disease resistance response gene 49
41.	Barnase	93.	Double stranded ribonuclease
42.	Barstar	94.	Ds transposable element
43.	Beta-hemoglobin	95.	Elongase
44.	B-glucuronidase	96.	EPSPS
45.	C1 knockout	97.	Ethylene forming enzyme knockout
46.	C1 regulatory gene	98.	Ethylene receptor protein
47.	C2 knockout	99.	Ethylene receptor protein knockout
48.	C3 knockout	100.	Fatty acid elongase
49.	Caffeate O-methylthransferase	101.	Fluorescent protein
50.	Caffeate O-methyltransferase knockout	102.	G glycoprotein
51.	Caffeoyl CoA O-methyltransferase knockout	103.	Galactanase

52.	Casein	104.	Galanthus nivalis agglutinin

Table 2 – Part 1.(continued) Non-limiting examples of transgenic genes & gene knockouts

105.	Genome-linked protein	157.	Omega 3 desaturease knockout
106.	Glucanase	158.	Omega 6 desaturase
107.	Glucanase knockout	159.	Omega 6 desaturase knockout
108.	Glucose oxidase	160.	O-methyltransferase
109.	Glutamate dehydrogenase	161.	Osmotin
110.	Glutamine binding protein	162.	Oxalate oxidase
111.	Glutamine synthetase	163.	Par locus
112.	Glutenin	164.	Pathogenesis protein 1a
113.	Glycerol-3-phosphate acetyl transferase	165.	Pectate lyase
114.	Glyphosate exidoreductase	166.	Pectin esterase
115.	Glyphosate oxidoreductase	167.	Pectin esterase knockout
116.	Green fluorescent protein	168.	Pectin methylesterase
117.	Helper component	169.	Pectin methylesterase knockout
118.	Hemicellulase	170.	Pentenlypyrophosphate isomerase
119.	Hup locus	171.	Phosphinothricin
120.	Hygromycin phosphotransferase	172.	Phosphinothricin acetyl transferase
121.	Hyoscamine 6B-hydroxylase	173.	Phytochrome A
122.	IAA monooxygenase	174.	Phytoene synthase
123.	Invertase	175.	Phleomycin binding protein
124.	Invertase knockout	176.	Polygalacturonase
125.	Isopentenyl transferase	177.	Polygalacturonase knockout
126.	Ketoacyl-ACP synthase	178.	Polygalacturonase inhibitor protein
127.	Ketoacyl-ACP synthase knockout	179.	Prf regulatory gene
128.	Larval serum protein	180.	Prosystemin
129.	Leafy homeotic regulatory gene	181.	Protease
130.	Lectin	182.	Protein A
131.	Lignin peroxidase	183.	Protein kinase
132.	Luciferase	184.	Proteinase inhibitor I
133.	Lysine-2 gene	185.	Pti5 transcription factor
134.	Lysophosphatidic acid acetyl transferase	186.	R regulatory gene
135.	Lysozyme	187.	Receptor kinase
136.	Mabinlin	188.	Recombinase
137.	Male sterility protein	189.	Reductase
138.	Metallothionein	190.	Replicase
139.	Modified ethylene receptor protein	191.	Resveratrol synthase
140.	Modified ethylene receptor protein knockout	192.	Ribonuclease
141.	Monooxygenase	193.	rolc
142.	Movement protein	194.	Rol hormone gene
143.	Movement protein nonfunctional	195.	S-adenosylmethione decarboxylase
144.	N gene for TMV resistance	196.	S-adenosylmethione hydrolase
145.	N-acetyl glucosidase	197.	S-adenosylmethionine transferase
146.	Nitrilase	198.	Salicylate hydroxylase
147.	Nopaline synthase	199.	Satellite RNA
148.	Notch	200.	Seed storage protein
149.	NptII	201.	Serine-threonine protein kinase
150.	Nuclear inclusion protein a	202.	Serum albumin
151.	Nuclear inclusion protein b	203.	Shrunken 2
152.	Nucleocapsid	204.	Sorbitol dehydrogenase
153.	Nucleoprotein	205.	Sorbitol synthase
154.	O-acyl transferase	206.	Stilbene synthase
155.	Oleayl-ACP thioesterase	207.	Storage protein
156.	Omega 3 desaturase	208.	Sucrose phosphate synthase

Table 2 – Part 1.(continued) Non-limiting examples of transgenic genes & gene knockouts

209.	Systemic acquired resistance gene 8.2	219.	Trichosanthin

210.	Tetracycline binding protein	220.	Trifolitoxin
211.	Thioesterase (x2)	221.	Trypsin inhibitor
212.	Thiolase	222.	T-URF13 mitochondrial
213.	TobRB7	223.	UDP glucose glucosyltransferase
214.	Transcriptional activator	224.	Violaxanthin de-epoxidase
215.	Transposon Tn5	225.	Violaxanthin de-epoxidase knockout
216.	Trehalase	226.	Wheat germ agglutinin
217.	Trehalase knockout	227.	Xanthosine-N7-methyltransferase knockout
218.	Trichodiene synthase	228.	Zein storage protein

Table 2 – Part 2. Non-limiting examples of input traits/phenotypes

1.	2,4-D tolerant	52.	Flowering time altered
2.	Alemaria resistant	53.	Frogeye leaf spot resistant
3.	Altered amino acid composition	54.	Fruit ripening altered
4.	Alternaria solani resistant	55.	Fruit ripening delayed
5.	Ammonium assimilation increased	56.	Fruit rot resistant
6.	AMV resistant	57.	Fruit solids increased
7.	Aphid resistant	58.	Fruit sweetness increased
8.	Apple scab resistant	59.	Fungal post-harvest resistant
9.	Aspergillus resistant	60.	Fungal resistant
10.	B-1,4-endoglucanase	61.	Fungal resistant general
11.	Bacterial leaf blight resistant	62.	Fusarium resistant
12.	Bacterial speck resistant	63.	Glyphosate tolerant
13.	BCTV resistant	64.	Growth rate altered
14.	Blackspot bruise resistant	65.	Growth rate reduced
15.	BLRV resistant	66.	Heat stable glucanase produced
16.	BNYVV Resistant	67.	Hordothionin produced
17.	Botrytis cinerea resistant	68.	Imidazolinone tolerant
18.	Botrytis resistant	69.	Insect resistant general
19.	BPMV resistant	70.	Kanamycin resistant
20.	Bromoxynil tolerant	71.	Lepidopteran resistant
21.	BYDV resistant	72.	Lesser cornstalk borer resistant
22.	BYMV resistant	73.	LMV resistant
23.	Carbohydrate metabolism altered	74.	Loss of systemic resistance
24.	Cell wall altered	75.	Male sterile
25.	Chlorsulfuron tolerant	76.	Marssonina resistant
26.	Clavibacter resistant	77.	MCDV resistant
27.	CLRV resistant	78.	MCMV resistant
28.	CMV resistant	79.	MDMV resistant
29.	Cold tolerant	80.	MDMV-B resistant
30.	Coleopteran resistant	81.	Mealybug wilt virus resistant
31.	Colletotrichum resistant	82.	Melamtsora resistant
32.	Colorado potato beetle resistant	83.	Melodgyne resistant
33.	Constitutive expression of glutamine synthetase	84.	Methotrexate resistant
34.	Corynebacterium sepedonicum resistant	85.	Mexican Rice Borer resistant
35.	Cottonwood leaf beetle resistant	86.	Nucleocapsid protein produced
36.	Crown gall resistant	87.	Oblique banded leafroller resistant
37.	Crown rot resistant	88.	PEMV resistant
38.	Cucumovirus resistant	89.	PeSV resistant
39.	Cutting rootability increased	90.	Phoma resistant
40.	Downy mildew resistant	91.	Phosphinothricin tolerant
41.	Drought tolerant	92.	Phratora leaf beetle resistant
42.	Erwinia carotovora resistant	93.	Phytophthora resistant
43.	Ethylene production reduced	94.	PLRV resistant
44.	European Corn Borer resistant	95.	Polyamine metabolism altered
45.	Female sterile	96.	Potyvirus resistant
46.	Fenthion susceptible	97.	Powdery mildew resistant
47.	Fertility altered	98.	PPV resistant
48.	Fire blight resistant	99.	Pratylenchus vulnus resistant

49.	Flower and fruit abscission reduced	100.	Proteinase inhibitors level constitutive
50.	Flower and fruit set altered	101.	PRSV resistant
51.	Flowering altered	102.	PRV resistant

Table 2 – Part 2.(continued)Non-limiting examples of transgenic input traits/phenotypes

103.	PSbMV resistant	128.	Streptomyces scabies resistant
104.	Pseudomonas syringae resistant	129.	Sulfonylurea tolerant
105.	PStV resistant	130.	Tetracycline binding protein produced
106.	PVX resistant	131.	TEV resistant
107.	PVY resistant	132.	Thelaviopsis resistant
108.	RBDV resistant	133.	TMV resistant
109.	Rhizoctonia resistant	134.	Tobamovirus resistant
110.	Rhizoctonia solani resistant	135.	ToMoV resistant
111.	Ring rot resistance	136.	ToMV resistant
112.	Root-knot nematode resistant	137.	Transposon activator
113.	SbMV resistant	138.	Transposon inserted
114.	Sclerotinia resistant	139.	TRV resistant
115.	SCMV resistant	140.	TSWV resistant
116.	SCYLV resistant	141.	TVMV resistant
117.	Secondary metabolite increased	142.	TYLCV resistant
118.	Seed set reduced	143.	Tyrosine level increased
119.	Selectable marker	144.	Venturia resistant
120.	Senescence altered	145.	Verticillium dahliae resistant
121.	Septoria resistant	146.	Verticillium resistant
122.	Shorter stems	147.	Visual marker
123.	Soft rot fungal resistant	148.	WMV2 resistant
124.	Soft rot resistant	149.	WSMV resistant
125.	SqMV resistant	150.	Yield increased
126.	SrMV resistant	151.	ZYMV resistant
127.	Storage protein altered		

Table 2 – Part 3. Non-limiting examples of output traits/phenotypes

1.	ACC oxidase level decreased	36.	Oil profile altered
2.	Altered lignin biosynthesis	37.	Pectin esterase level reduced
3.	B-1,4-endoglucanase	38.	Pharmaceutical proteins produced
4.	Botrytis resistant	39.	Phosphinothricin tolerant
5.	Carbohydrate metabolism altered	40.	Phytoene synthase activity increased
6.	Carotenoid content altered	41.	Pigment metabolism altered
7.	Cell wall altered	42.	Polygalacturonase level reduced
8.	CMV resistant	43.	Processing characteristics altered
9.	Coleopteran resistant	44.	Prolonged shelf life
10.	Dry matter content increased	45.	Protein altered
11.	Ethylene production reduced	46.	Protein quality altered
12.	Ethylene synthesis reduced	47.	PRSV resistant
13.	Fatty acid metabolism altered	48.	Root-knot nematode resistant
14.	Fire blight resistant	49.	Sclerotinia resistant
15.	Flower and fruit abscission reduced	50.	Seed composition altered
16.	Flower and fruit set altered	51.	Seed methionine storage increased
17.	Flowering time altered	52.	Seed set reduced
18.	Fruit firmness increased	53.	Seed storage protein
19.	Fruit pectin esterase levels decreased	54.	Senescence altered (e.g. Shelf life increased)
20.	Fruit ripening altered	55.	Shorter stems
21.	Fruit ripening delayed	56.	Solids increased
22.	Fruit solids increased	57.	SqMV resistant
23.	Fruit sugar profile altered	58.	Starch level increased
24.	Fruit sweetness increased	59.	Starch metabolism altered
25.	Glucuronidase expressing	60.	Starch reduced
26.	Heat stable glucanase produced	61.	Sterols increased
27.	Heavy metals sequestered	62.	Storage protein altered

28. Hordothionin produced	63. Sugar alcohol levels increased
29. Improved fruit quality	64. Tetracycline binding protein produced
30. Industrial enzyme produced	65. Tyrosine level increased
31. Lepidopteran resistant	66. Verticillium resistant
32. Lysine level increased	67. Visual marker
33. Mealybug wilt virus resistant	68. WMV2 resistant
34. Methionine level increased	69. Yield increased
35. Nucleocapsid protein produced	70. ZYMV resistant

Table 2 – Part 4. Non-limiting examples of traits/phenotypes with agronomic properties

1. ACC oxidase level decreased	53. Industrial enzyme produced
2. Altered amino acid composition	54. Lignin levels decreased
3. Altered lignin biosynthesis	55. Lipase expressed in seeds
4. Altered maturing	56. Lysine level increased
5. Altered plant development	57. Male sterile
6. Aluminum tolerant	58. Male sterile reversible
7. Ammonium assimilation increased	59. Methionine level increased
8. Anthocyanin produced in seed	60. Modified growth characteristics
9. B-1,4-endoglucanase	61. Mycotoxin degradation
10. Calmodulin level altered	62. Nitrogen metabolism altered
11. Carbohydrate metabolism altered	63. Nucleocapsid protein produced
12. Carotenoid content altered	64. Oil profile altered
13. Cell wall altered	65. Oil quality altered
14. Cold tolerant	66. Oxidative stress tolerant
15. Constitutive expression of glutamine synthetase	67. Pectin esterase level reduced
16. Cutting root ability increased	68. Pharmaceutical proteins produced
17. Development altered	69. Photosynthesis enhanced
18. Drought tolerant	70. Phytoene synthase activity increased
19. Dry matter content increased	71. Pigment metabolism altered
20. Environmental stress reduced	72. Polyamine metabolism altered
21. Ethylene metabolism altered	73. Polygalacturonase level reduced
22. Ethylene production reduced	74. Pratylenchus vulnus resistant
23. Ethylene synthesis reduced	75. Processing characteristics altered
24. Fatty acid metabolism altered	76. Prolonged shelf life
25. Female sterile	77. Protein altered
26. Fenthion susceptible	78. Protein lysine level increased
27. Fertility altered	79. Protein quality altered
28. Fiber quality altered	80. Proteinase inhibitors level constitutive
29. Flower and fruit abscission reduced	81. Salt tolerance increased
30. Flower and fruit set altered	82. Seed composition altered
31. Flowering altered	83. Seed methionine storage increased
32. Flower color altered	84. Seed set reduced
33. Flowering time altered	85. Selectable marker
34. Fruit firmness increased	86. Senescence altered
35. Fruit pectin esterase and levels decreased	87. Shorter stems
36. Fruit polygalacturonase level decreased	88. Solids increased
37. Fruit ripening altered	89. Starch level increased
38. Fruit ripening delayed	90. Starch metabolism altered
39. Fruit solids increased	91. Starch reduced
40. Fruit sugar profile altered	92. Sterols increased
41. Fruit sweetness increased	93. Storage protein altered
42. Glucuronidase expressing	94. Stress tolerant
43. Growth rate altered	95. Sugar alcohol levels increased
44. Growth rate increased	96. Tetracycline binding protein produced
45. Growth rate reduced	97. Thermostable protein produced
46. Heat stable glucanase produced	98. Transposon activator
47. Heat tolerant	99. Transposon inserted
48. Heavy metals sequestered	100. Tyrosine level increased
49. Hordothionin produced	101. Visual marker
50. Improved fruit quality	102. Vivipary increased

51.	Increased phosphorus	103.	Yield increased
52.	Increased stalk strength		

Table 2 – Part 5. Non-limiting examples of traits/phenotypes with product quality properties

1.	2,4-D tolerant	45.	Melanin produced in cotton fibers
2.	ACC oxidase level decreased	46.	Metabolism altered
3.	Altered amino acid composition	47.	Methionine level increased
4.	Altered lignin biosynthesis	48.	Mycotoxin degradation
5.	Anthocyanin produced in seed	49.	Mycotoxin production inhibited
6.	Antioxidant enzyme increased	50.	Nicotine levels reduced
7.	Auxin metabolism and increased tuber solids	51.	Nitrogen metabolism altered
8.	B-1,4-endoglucanase	52.	Novel protein produced
9.	Blackspot bruise resistant	53.	Nutritional quality altered
10.	Brown spot resistant	54.	Oil profile altered
11.	Bruising reduced	55.	Oil quality altered
12.	Caffeine levels reduced	56.	Pectin esterase level reduced
13.	Carbohydrate metabolism altered	57.	Photosynthesis enhanced
14.	Carotenoid content altered	58.	Phytoene synthase activity increased
15.	Cell wall altered	59.	Pigment metabolism altered
16.	Cold tolerant	60.	Polyamine metabolism altered
17.	Delayed softening	61.	Polygalacturonase level reduced
18.	Disulfides reduced in endosperm	62.	Processing characteristics altered
19.	Dry matter content increased	63.	Prolonged shelf life
20.	Ear mold resistant	64.	Protein altered
21.	Ethylene production reduced	65.	Protein lysine level increased
22.	Ethylene synthesis reduced	66.	Protein quality altered
23.	Extended flower life	67.	Proteinase inhibitors level constitutive
24.	Fatty acid metabolism altered	68.	Rust resistant
25.	Fiber quality altered	69.	Seed composition altered
26.	Fiber strength altered	70.	Seed methionine storage increased
27.	Flavor enhancer	71.	Seed number increased
28.	Flower and fruit abscission reduced	72.	Seed quality altered
29.	Fruit firmness increased	73.	Seed set reduced
30.	Fruit invertase level decreased	74.	Seed weight increased
31.	Fruit polygalacturonase level decreased	75.	Senescence altered
32.	Fruit ripening altered	76.	Solids increased
33.	Fruit ripening delayed	77.	Starch level increased
34.	Fruit solids increased	78.	Starch metabolism altered
35.	Fruit sugar profile altered	79.	Starch reduced
36.	Fruit sweetness increased	80.	Steroidal glycoalkaloids reduced
37.	Glyphosate tolerant	81.	Sterols increased
38.	Heat stable glucanase produced	82.	Storage protein altered
39.	Improved fruit quality	83.	Sugar alcohol levels increased
40.	Increased phosphorus	84.	Thermostable protein produced
41.	Increased protein levels	85.	Tryptophan level increased
42.	Lignin levels decreased	86.	Tuber solids increased
43.	Lysine level increased	87.	Yield increased
44.	Male sterile		

Table 2 – Part 6. Non-limiting examples of traits/phenotypes with herbicide tolerance properties

1.	2,4-D tolerant	11.	Sulfonylurea tolerant
2.	Chloroacetanilide tolerant	12.	Northern corn leaf blight resistant
3.	Fertility altered	13.	Herbicide tolerant
4.	Protein altered	14.	Isoxazole tolerant
5.	Lignin levels decreased	15.	Chlorsulfuron tolerant
6.	Methionine level increased	16.	Glyphosate tolerant
7.	Bromoxynil tolerant	17.	Lepidopteran resistant
8.	Metabolism altered	18.	Phosphinothricin tolerant
9.	Imidazole tolerant	19.	Sulfonylurea tolerant

10.	Imidazolinone tolerant

Table 2 – Part 7. Non-limiting examples of traits/phenotypes with pest resistance properties

Legend

BR - Bacterial Resistant	**NR** - Nematode Resistant
FR - Fungal Resistant	VR - **Viral Resistant**
IR - Insent Resistant	

1.	Agrobacterium resistant – **BR**	44.	Ear mold resistant – **FR**
2.	Alternaria resistant – **FR**	45.	Erwinia carotovora resistant – **BR**
3.	Alternaria daucii resistant – **FR**	46.	European Corn Borer resistant – **IR**
4.	Alternaria solani resistant – **FR**	47.	Eyespot resistant – **FR**
5.	AMV resistant – **VR**	48.	Fall armyworm resistant – **IR**
6.	Anthracnose resistant – **FR**	49.	Fire blight resistant – **BR**
7.	Aphid resistant – **IR**	50.	Frogeye leaf spot resistanT – **FR**
8.	Apple scab resistant – **FR**	51.	Fruit rot resistant – **FR**
9.	Aspergillus resistant – **FR**	52.	Fungal post-harvest resistant – **FR**
10.	Bacterial leaf blight resistant – **BR**	53.	Fungal resistant – **FR**
11.	Bacterial resistant – **BR**	54.	Fungal resistant general – **FR**
12.	Bacterial soft rot resistant – **BR**	55.	Fusarium dehlae resistant – **FR**
13.	Bacterial soft rot resistant – **VR**	56.	Fusarium resistant – **FR**
14.	Bacterial speck resistant – **BR**	57.	Geminivirus resistant – **VR**
15.	BCTV resistant – **VR**	58.	Gray lead spot resistant – **FR**
16.	Black shank resistant – **FR**	59.	Helminthosporium resistant – **FR**
17.	BLRV resistant – **VR**	60.	Hordothionin produced – **BR**
18.	BNYVV resistant – **VR**	61.	Insect predator resistant – **IR**
19.	Botrytis cinerea resistant – **FR**	62.	Insect resistant general – **IR**
20.	Botrytis resistant – **FR**	63.	Late blight resistant – **FR**
21.	BPMV resistant – **VR**	64.	Leaf blight resistant – **FR**
22.	Brown spot resistant – **FR**	65.	Leaf spot resistant – **FR**
23.	BYDV resistant – **VR**	66.	Lepidopteran resistant – **IR**
24.	BYMV resistant – **VR**	67.	Lesser cornstalk borer resistant – **IR**
25.	CaMV resistant – **VR**	68.	LMV resistant – **VR**
26.	Cercospora resistant – **FR**	69.	Loss of systemic resistance – **VR**
27.	Clavibacter resistant – **BR**	70.	Marssonina resistant – **FR**
28.	Closteroviurs resistant – **BR**	71.	MCDV resistant – **VR**
29.	CLRV resistant – **VR**	72.	MCMV resistant – **VR**
30.	CMV resistant – **FR**	73.	MDMV resistant – **VR**
31.	Coleopteran resistant – **IR**	74.	MDMV-B resistant – **VR**
32.	Colletotrichum resistant – **FR**	75.	Mealybug wilt virus resistant – **VR**
33.	Colorado potato beetle resistant – **IR**	76.	Melamtsora resistant – **FR**
34.	Corn earworm resistant – **IR**	77.	Melodgyne resistant – **NR**
35.	Corynebacterium sepedonicum resistant – **BR**	78.	Meloidogyne resistant – **NR**
36.	Cottonwood leaf beetle resistant – **IR**	79.	Mexican Rice Borer resistant – **IR**
37.	Criconnemella resistant – **NR**	80.	Mycotoxin degradation – **FR**
38.	Crown gall resistant – **BR**	81.	Nepovirus resistant – **VR**
39.	Cucumovirus resistant – **VR**	82.	Northern corn leaf blight resistant – **IR**
40.	Cylindrosporium resistant – **FR**	83.	Nucleocapsid protein produced – **VR**
41.	Disease resistant general – **FR**	84.	Oblique banded leafroller resistant – **IR**
42.	Dollar spot resistant – **FR**	85.	Oomycete resistant – **FR**
43.	Downy mildew resistant – **FR**	86.	Pathogenesis related proteins level increased – **FR**

Table 2 – Part 7. (continued) Non-limiting examples of traits/phenotypes with pest resistance properties

87.	PEMV resistant – **VR**	116.	SMV resistant – **VR**
88.	PeSV Resistant – **VR**	117.	Sod web worm resistant – **IR**
89.	Phatora leaf beetle resistant – **IR**	118.	Soft rot fungal resistant – **FR**
90.	Phoma resistant – **FR**	119.	Soft rot resistant – **BR**

91.	Phytophthora resistant – **FR**	120.	Southwestern corn borer resistant – **IR**
92.	PLRV resistant – **VR**	121.	SPFMV resistant – **VR**
93.	Potyvirus resistant – **VR**	122.	Sphaeropsis fruit rot resistant – **FR**
94.	Powdery mildew resistant – **FR**	123.	SqMV resistant – **VR**
95.	PPV resistant – **VR**	124.	SrMV resistant – **VR**
96.	Pratylenchus vulnus resistant – **NR**	125.	Streptomyces scabies resistant – **BR**
97.	PRSV resistant – **VR**	126.	Sugar cane borer resistant – **IR**
98.	PRV resistant – **VR**	127.	TEV resistant – **VR**
99.	PSbMV resistant – **VR**	128.	Thelaviopsis resistant – **FR**
100.	Pseudomonas syringae resistant – **BR**	129.	TMV resistant – **FR**
101.	PStV resistant – **VR**	130.	Tobamovirus resistant – **VR**
102.	PVX resistant – **VR**	131.	ToMoV resistant – **VR**
103.	PVY resistant – **VR**	132.	ToMV resistant – **VR**
104.	RBDV resistant – **VR**	133.	TRV resistant – **VR**
105.	Rhizoctonia resistant – **FR**	134.	TSWV resistant – **VR**
106.	Rhizoctonia solani resistant – **FR**	135.	TVMV resistant – **VR**
107.	Ring rot resistance – **BR**	136.	TYLCV resistant – **VR**
108.	Root-knot nematode resistant – **NR**	137.	Venturia resistant – **FR**
109.	Rust resistant – **FR**	138.	Verticillium dahliae resistant – **FR**
110.	SbMV resistant – **VR**	139.	Verticillium resistant – **FR**
111.	Sclerotinia resistant – **FR**	140.	Western corn root worm resistant – **IR**
112.	SCMV resistant – **VR**	141.	WMV2 resistant – **VR**
113.	SCYLV resistant – **VR**	142.	WSMV resistant – **VR**
114.	Septoria resistant – **FR**	143.	ZYMV resistant – **VR**
115.	Smut resistant – **FR**		

Table 2 – Part 8. Non-limiting examples of miscellaneous traits/phenotypes with properties

1.	Antibiotic produced	31.	Mycotoxin production inhibited
2.	Antiprotease producing	32.	Mycotoxin restored
3.	Capable of growth on defined synthetic media	33.	Non-lesion forming mutant
4.	Carbohydrate metabolism altered	34.	Novel protein produced
5.	Cell wall altered	35.	Oil quality altered
6.	Cold tolerant	36.	Peroxidase levels increased
7.	Coleopteran resistant	37.	Pharmaceutical proteins produced
8.	Color altered	38.	Phosphinothricin tolerant
9.	Color sectors in seeds	39.	Pigment metabolism altered
10.	Colored sectors in leaves	40.	Pollen visual marker
11.	Constitutive expression of glutamine synthetase	41.	Polyamine metablosim altered
12.	Cre recombinase produced	42.	Polymer produced
13.	Dalapon tolerant	43.	Recombinase produced
14.	Development altered	44.	Secondary metabolite increased
15.	Disease resistant general	45.	Seed color altered
16.	Ethylene metabolism altered	46.	Seed weight increased
17.	Expression optimization	47.	Selectable marker
18.	Fenthion susceptible	48.	Spectromycin resistant
19.	Glucuronidase expressing	49.	Sterile
20.	Glyphosate tolerant	50.	Sterols increased
21.	Growth rate reduced	51.	Sulfonylurea susceptible
22.	Heavy metals sequestered	52.	Syringomycin deficient
23.	Hygromycin tolerant	53.	Transposon activator
24.	Inducible DNA modification	54.	Transposon elements inserted
25.	Industrial enzyme produced	55.	Transposon inserted
26.	Kanamycin resistant	56.	Trifolitoxin producing
27.	Lipase expressed in seeds	57.	Trifolitoxin resistant
28.	Methotrexate resistant	58.	Virulence reduced
29.	Modified growth characteristics	59.	Visual marker
30.	Mycotoxin deficient	60.	Visual marker inactive

In a particular examplification, “**producing an organism having a desirable trait**” includes an organism that is with respect to an organ or a part of an organ but not necessarily altered anywhere else.

By "**trait**" is meant any detectable parameter associated with an organism under a set of conditions. Examples of "detectable parameters" include the ability to produce a substance, the ability to not produce a substance, an altered pattern of (such as an increased or a decreased) ability to produce a substance, viability, non-viability, behaviour, growth rate, size, morphology or morphological characteristic,

In another embodiment, this invention is directed to a method of producing an organism having a desirable trait or a desirable improvement in a trait by: a) obtaining an initial population of organisms comprised of at least one starting organism, b) mutagenizing the population such that mutations occur throughout a substantial part of the genome of at least one initial organism, c) selecting at least one mutagenized organism having a desirable trait or a desirable improvement in a trait, and d) optionally repeating the method by subjecting one or more mutagenized organisms to a repetition of the method. A mutagenized organism having a desirable trait or a desirable improvement in a trait can be referred to as an "up-mutant", and the associated mutation(s) contained in an up-mutant organism can be referred to as up-mutation(s).

In one embodiment, step c) is comprised of selecting at least two different mutagenized organisms, each having a different mutagenized genome, and the method of producing an organism having a desirable trait or a desirable improvement in a trait is comprised of a) obtaining a starting population of organisms comprised of at least one starting organism, b) mutagenizing the population such that mutations occur throughout a substantial part of the genome of at least one starting organism, c) selecting at least two mutagenized organism having a desirable trait or a desirable improvement in a trait, d) creating combinations of the mutations of the two or more mutagenized organisms, e) selecting at least one mutagenized organism having a desirable trait or a desirable improvement in a trait, and f) optionally repeating the method by subjecting one or more mutagenized organisms to a repetition of the method.

In one embodiment, the method is repeated. Thus, for example, an up-mutant organism can serve as a starting organism for the above method. Also, for example, an up mutant organism having a combination of two or more up-mutations in its genome can serve as a starting organism for the above method.

Thus, in one embodiment, this invention is directed to a method of producing an organism having a desirable trait or a desirable improvement in a trait by: a) obtaining a starting population of organisms comprised of at least one starting organism, b) mutagenizing the population such that mutations occur throughout a substantial part of the genome of at least one starting organism, c) selecting at least one mutagenized organism having a desirable trait or a desirable improvement in a trait, and d) optionally repeating the method by subjecting one or more mutagenized organisms to a repetition of the method. A mutagenized organism having a desirable trait or a desirable improvement in a trait can be referred to as an "up-mutant", and the associated mutation(s) contained in an up-mutant organism can be referred to as up-mutation(s).

Mutagenizing a starting population such that mutations occur throughout a substantial part of the genome of at least one starting organism refers to mutagenizing at least approximately 1% of the genes of a genome, or at least approximately 10% of the genes of a genome, or at least approximately 20% of the genes of a genome, or at least approximately 30% of the genes of a genome, or at least approximately 40% of the genes of a genome, or at least approximately 50% of the genes of a genome, or at least approximately 60% of the genes of a genome, or at least approximately 70% of the genes of a genome, or at least approximately 80% of the genes of a genome, or at least approximately 90% of the genes of a genome, or at least approximately 95% of the genes of a genome, or at least approximately 98% of the genes of a genome.

In a particular embodiment, this invention provides a method of producing an organism having a desirable trait or a desirable improvement in a trait by: a) obtaining sequence information of a genome; b) annotating the genomic sequence obtained; c) mutagenizing a substantial part of the genome the genome; d) selecting at least one mutagenized genome having a desirable trait or a desirable improvement in a trait; and e) optionally repeating the method by subjecting one or more mutagenized genomes to a repetition of the method.

Thus in one aspect, this invention provides a process comprised of:

1.) Subjecting a working cell or organism to holistic monitoring (which can include the detection and/or measurement of all detectable functions and physical parameters). Examples of such parameters include morphology, behavior, growth, responsiveness to stimuli (e.g., antibiotics, different environment, etc.). Additional examples include all measurable molecules, including molecules that are chemically at least in part a nucleic acids, proteins, carbohydrates, proteoglycans, glycoproteins, or lipids. In a particular aspect, performing holistic monitoring is comprised of using a microarray-based method. In another aspect, performing holistic monitoring is comprised of sequencing a substantial portion of the genome, i.e. for example at least approximately 10% of the genome, or for example at least approximately 20% of the genome, or for example at least approximately 30% of the genome, or for example at least approximately 40% of the genome, or for example at least approximately 50% of the genome, or for example at least approximately 60% of the genome, or for example at least approximately 70% of the genome, or for example at least approximately 80% of the genome, or for example at least approximately 90% of the genome, or for example at least approximately 95% of the genome, or for example at least approximately 98% of the genome.

2) Introducing into the working cell or organism a plurality of traits (stacked traits), including selectively and differentially activatable traits. Serviceable traits for this purpose include traits conferred by genes and traits conferred by gene pathways.

3) Subjecting the working cell or organism to holistic monitoring.

4) Compiling the information obtained from steps 1) and 3), and processing &/or analyzing it to better understand the changes introduced into the working cell or organisms. Such data processing includes identifying correlations between and/or among the measured parameters.

5) Repeating any number or all of steps 2), 3), and 4).

This invention provides that molecules serviceable for introducing transgenic traits into a plant include all known genes and nucleic acids. By way of non-limiting exemplification, this invention specifically names any number &/or combination of genes listed herein or listed in any reference incorporated herein by reference . Furthermore, by way of non-limiting exemplification, this invention specifically names any number &/or combination of genes & gene pathways listed herein as well as in any reference incorporated by reference herein. This invention provides that molecules serviceable as detectable parameters include molecule, any enzyme, substrate thereof, product thereof, and any gene or gene pathway listed herein including in any figure or table herein as well as in any reference incorporated by reference herein.

This invention also relates generally to the field of nucleic acid engineering and correspondingly encoded recombinant protein engineering. More particularly, the invention relates to the directed evolution of nucleic acids and screening of clones containing the evolved nucleic acids for resultant activity(ies) of interest, such nucleic acid activity(ies) &/or specified protein, particularly enzyme, activity(ies) of interest.

Mutagenized molecules provided by this invention may have chimeric molecules and molecules with point mutations, including biological molecules that contain a carbohydrate, a lipid, a nucleic acid, &/or a protein component, and specific but non-limiting examples of these include antibiotics, antibodies, enzymes, and steroidal and non-steroidal hormones.

This invention relates generally to a method of: 1) preparing a progeny generation of molecule(s) (including a molecule that is comprised of a polynucleotide sequence, a molecule that is comprised of a polypeptide sequence, and a molecules that is comprised in part of a polynucleotide sequence and in part of a polypeptide sequence), that is mutagenized to achieve at least one point mutation, addition, deletion, &/or chimerization, from one or more ancestral or parental generation template(s); 2) screening the progeny generation molecule(s) - preferably using a high throughput method - for at least one property of interest (such as an improvement in an enzyme activity or an increase in stability or a novel chemotherapeutic effect); 3) optionally obtaining &/or cataloguing structural &/or and functional information regarding the parental &/or progeny generation molecules; and 4) optionally repeating any of steps 1) to 3).

In a preferred embodiment, there is generated (e.g. from a parent polynucleotide template) - in what is termed "codon site-saturation mutagenesis" - a progeny generation of polynucleotides, each having at least one set of up to three contiguous point mutations (i.e. different bases comprising a new codon), such that every codon (or every family of degenerate codons encoding the same amino acid) is represented at each codon position. Corresponding to - and encoded by - this progeny generation of polynucleotides, there is also generated a set of progeny polypeptides, each having at least one single amino acid point mutation. In a preferred aspect, there is generated - in what is termed "amino acid site-saturation mutagenesis" - one such mutant polypeptide for each of the 19 naturally encoded polypeptide-forming alpha-amino acid substitutions at each and every amino acid position along the polypeptide. This yields - for each and every amino acid position along the parental polypeptide - a total of 20 distinct progeny polypeptides including the original amino acid, or potentially more than 21 distinct progeny polypeptides if additional amino acids are used either instead of or in addition to the 20 naturally encoded amino acids

Thus, in another aspect, this approach is also serviceable for generating mutants containing - in addition to &/or in combination with the 20 naturally encoded polypeptide-forming alpha-amino acids - other rare &/or not naturally-encoded amino acids and amino acid derivatives. In yet another aspect, this approach is also serviceable for generating mutants by the use of - in addition to &/or in combination with natural or unaltered codon recognition systems of suitable hosts - altered, mutagenized, &/or designer codon recognition systems (such as in a host cell with one or more altered tRNA molecules).

In yet another aspect, this invention relates to recombination and more specifically to a method for preparing polynucleotides encoding a polypeptide by a method of *in vivo* re-assortment of polynucleotide sequences containing regions of partial homology, assembling the polynucleotides to form at least one polynucleotide and screening the polynucleotides for the production of polypeptide(s) having a useful property.

In yet another preferred embodiment, this invention is serviceable for analyzing and cataloguing - with respect to any molecular property (e.g. an enzymatic activity) or combination of properties allowed by current technology - the effects of any mutational change achieved (including particularly saturation mutagenesis). Thus, a comprehensive method is provided for determining the effect of changing each amino acid in a parental polypeptide into each of at least 19 possible substitutions. This allows each amino acid in a parental polypeptide to be characterized and catalogued according to its spectrum of potential effects on a measurable property of the polypeptide.

In another aspect, the method of the present invention utilizes the natural property of cells to recombine molecules and/or to mediate reductive processes that reduce the complexity of sequences and extent of repeated or consecutive sequences possessing regions of homology.

It is an object of the present invention to provide a method for generating hybrid polynucleotides encoding biologically active hybrid polypeptides with enhanced activities. In accomplishing these and other objects, there has been provided, in accordance with one aspect of the invention, a method for introducing polynucleotides into a suitable host cell and growing the host cell under conditions that produce a hybrid polynucleotide.

In another aspect of the invention, the invention provides a method for screening for biologically active hybrid polypeptides encoded by hybrid polynucleotides. The present method allows for the identification of biologically active hybrid polypeptides with enhanced biological activities.

Other objects, features and advantages of the present invention will become apparent from the following detailed description. It should be understood, however, that the detailed description and the specific examples, while indicating preferred embodiments of the invention, are given by way of illustration only, since various changes and modifications within the spirit and scope of the invention will become apparent to those skilled in the art from this detailed description.

In yet another aspect, this invention relates to a method of discovering which phenotype corresponds to a gene by disrupting every gene in the organism.

Accordingly, this invention provides a method for determining a gene that alters a characteristic of an organism, comprising: a) obtaining an initial population of organisms, b) generating a set of mutagenized organisms, such that when all the genetic mutations in the set of mutagenized organisms are taken as a whole, there is represented a set of substantial genetic mutations, and c) detecting the presence an organism having an altered trait, and d) determining the nucleotide sequence of a gene that has been mutagenized in the organism having the altered trait.

In yet another aspect, this invention relates to a method of improving a trait in an organism by functionally knocking out a particular gene in the organism, and then transferring a library of genes, which only vary from the wild-type at one codon position, into the organism.

Accordingly, this invention provides a method method for producing an organism with an improved trait, comprising:

a) functionally knocking out an enogenous gene in a substantially clonal population of organisms;

b) transferring the set of altered genes into the clonal population of organisms, wherein each altered gene differs from the endogenous gene at only one codon; and

c) detecting a mutagenized organism having an improved trait; and

d) determining the nucleotide sequence of a gene that has been transferred into the detected organism.

D. BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1. Exonuclease Activity. Figure 1 shows the activity of the enzyme exonuclease III. This is an exemplary enzyme that can be used to shuffle, assemble, reassemble, recombine, and/or concatenate polynucleotide building blocks. The asterisk indicates that the enzyme acts from the 3' direction towards the 5' direction of the polynucleotide substrate.

Figure 2. Generation of A Nucleic Acid Building Block by Polymerase-Based Amplification. Figure 2 illustrates a method of generating a double-stranded nucleic acid

building block with two overhangs using a polymerase-based amplification reaction (e.g., PCR). As illustrated, a first polymerase-based amplification reaction using a first set of primers, F_2 and R_1, is used to generate a blunt-ended product (labeled Reaction 1, Product 1), which is essentially identical to Product A. A second polymerase-based amplification reaction using a second set of primers, F_1 and R_2, is used to generate a blunt-ended product (labeled Reaction 2, Product 2), which is essentially identical to Product B. These two products are then mixed and allowed to melt and anneal, generating a potentially useful double-stranded nucleic acid building block with two overhangs. In the example of Fig. 1, the product with the 3' overhangs (Product C) is selected for by nuclease-based degradation of the other 3 products using a 3' acting exonuclease, such as exonuclease III. Alternate primers are shown in parenthesis to illustrate serviceable primers may overlap, and additionally that serviceable primers may be of different lengths, as shown.

FIGURE 3. Unique Overhangs And Unique Couplings. Figure 3 illustrates the point that the number of unique overhangs of each size (e.g. the total number of unique overhangs composed of 1 or 2 or 3, etc. nucleotides) exceeds the number of unique couplings that can result from the use of all the unique overhangs of that size. For example, there are 4 unique 3' overhangs composed of a single nucleotide, and 4 unique 5' overhangs composed of a single nucleotide. Yet the total number of unique couplings that can be made using all the 8 unique single-nucleotide 3' overhangs and single-nucleotide 5' overhangs is 4.

FIGURE 4. Unique Overall Assembly Order Achieved by Sequentially Coupling the Building Blocks

Figure 4 illustrates the fact that in order to assemble a total of "n" nucleic acid building blocks, "n-1" couplings are needed. Yet it is sometimes the case that the number of unique couplings available for use is fewer that the "n-1" value. Under these, and other, circumstances a stringent non-stochastic overall assembly order can still be achieved by performing the assembly process in sequential steps. In this example, 2 sequential steps are used to achieve a designed overall assembly order for five nucleic acid building blocks. In this illustration the designed overall assembly order for the five nucleic acid building blocks is: 5'-(#1-#2-#3-#4-#5)-3', where #1 represents building block number 1, etc.

FIGURE 5. Unique Couplings Available Using a Two-Nucleotide 3' Overhang. Figure 5 further illustrates the point that the number of unique overhangs of each size (here, e.g. the total number of unique overhangs composed of 2 nucleotides) exceeds the number of unique couplings that can result from the use of all the unique overhangs of that size. For example, there are 16 unique 3' overhangs composed of two nucleotides, and another 16 unique 5' overhangs composed of two nucleotides, for a total of 32 as shown. Yet the total number of couplings that are unique and not self-binding that can be made using all the 32 unique double-nucleotide 3' overhangs and double-nucleotide 5' overhangs is 12. Some apparently unique couplings have "identical twins" (marked in the same shading), which are visually obvious in this illustration. Still other overhangs contain nucleotide sequences that can self-bind in a palindromic fashion, as shown and labeled in this figure; thus they not contribute the high stringency to the overall assembly order.

Figure 6. Generation of an Exhaustive Set of Chimeric Combinations by Synthetic Ligation Reassembly. Figure 6 showcases the power of this invention in its ability to generate exhaustively and systematically all possible combinations of the nucleic acid building blocks designed in this example. Particularly large sets (or libraries) of progeny chimeric molecules can be generated. Because this method can be performed exhaustively and systematically, the method application can be repeated by choosing new demarcation points and with correspondingly newly designed nucleic acid building blocks, bypassing the burden of re-generating and re-screening previously examined and rejected molecular species. It is appreciated that, codon wobble can be used to advantage to increase the frequency of a demarcation point. In other words, a particular base can often be substituted into a nucleic acid building block without altering the amino acid encoded by progenitor codon (that is now altered codon) because of codon degeneracy. As illustrated, demarcation points are chosen upon alignment of 8 progenitor templates. Nucleic acid building blocks including their overhangs (which are serviceable for the formation of ordered couplings) are then designed and synthesized. In this instance, 18 nucleic acid building blocks are generated based on the sequence of each of the 8 progenitor templates, for a total of 144 nucleic acid building blocks (or double-stranded oligos). Performing the

ligation synthesis procedure will then produce a library of progeny molecules comprised of yield of 8^{18} (or over 1.8×10^{16}) chimeras.

Figure 7. Synthetic genes from oligos:. According to one embodiment of this invention, double-stranded nucleic acid building blocks are designed by aligning a plurality of progenitor nucleic acid templates. Preferably these templates contain some homology and some heterology. The nucleic acids may encode related proteins, such as related enzymes, which relationship may be based on function or structure or both. Figure 7 shows the alignment of three polynucleotide progenitor templates and the selection of demarcation points (boxed) shared by all the progenitor molecules. In this particular example, the nucleic acid building blocks derived from each of the progenitor templates were chosen to be approximately 30 to 50 nucleotides in length.

Figure 8. Nucleic acid building blocks for synthetic ligation gene reassembly. Figure 8 shows the nucleic acid building blocks from the example in Figure 7. The nucleic acid building blocks are shown here in generic cartoon form, with their compatible overhangs, including both 5' and 3' overhangs. There are 22 total nucleic acid building blocks derived from each of the 3 progenitor templates. Thus, the ligation synthesis procedure can produce a library of progeny molecules comprised of yield of 3^{22} (or over 3.1×10^{10}) chimeras.

Figure 9. Addition of Introns by Synthetic Ligation Reassembly. Figure 9 shows in generic cartoon form that an intron may be introduced into a chimeric progeny molecule by way of a nucleic acid building block. It is appreciated that introns often have consensus sequences at both termini in order to render them operational. It is also appreciated that, in addition to enabling gene splicing, introns may serve an additional purpose by providing sites of homology to other nucleic acids to enable homologous recombination. For this purpose, and potentially others, it may be sometimes desirable to generate a large nucleic acid building block for introducing an intron. If the size is overly large easily genrating by direct chemical synthesis of two single stranded oligos, such a specialized nucleic acid building block may also be generated by direct chemical synthesis of more than two single stranded oligos or by using a polymerase-based amplification reaction as shown in Figure 2.

Figure 10. Ligation Reassembly Using Fewer Than All The Nucleotides Of An Overhang. Figure 10 shows that coupling can occur in a manner that does not make use of every nucleotide in a participating overhang. The coupling is particularly lively to survive (e.g. in a transformed host) if the coupling reinforced by treatment with a ligase enzyme to form what may be referred to as a "gap ligation" or a "gapped ligation". It is appreciated that, as shown, this type of coupling can contribute to generation of unwanted background product(s), but it can also be used advantageously increase the diversity of the progeny library generated by the designed ligation reassembly.

Figure 11. Avoidance of unwanted self-ligation in palindromic couplings. As mentioned before and shown in Figure 5, certain overhangs are able to undergo self-coupling to form a palindromic coupling. A coupling is strengthened substantially if it is reinforced by treatment with a ligase enzyme. Accordingly, it is appreciated that the lack of 5' phosphates on these overhangs, as shown, can be used advantageously to prevent this type of palindromic self-ligation. Accordingly, this invention provides that nucleic acid building blocks can be chemically made (or ordered) that lack a 5' phosphate group (or alternatively they can be remove – e.g. by treatment with a phosphatase enzyme such as a calf intestinal alkaline phosphatase (CIAP) – in order to prevent palindromic self-ligations in ligation reassembly processes.

Figure 12. Pathway Engineering. It is a goal of this invention to provide ways of making new gene pathways using ligation reassembly, optionally with other directed evolution methods such as saturation mutagenesis. Figure 12 illustrates a preferred approach that may be taken to achieve this goal. It is appreciated that naturally-occurring microbial gene pathways are linked more often than naturally-occurring eukaryotic (e.g. plant) gene pathways, which are sometime only partially linked. In a particular embodiment, this invention provides that regulatory gene sequences (including promoters) can be introduced in the form of nucleic acid building blocks into progeny gene pathways generated by ligation reassembly processes. Thus, originally linked microbial gene pathways, as well as originally unlinked genes and gene pathways, can be thus converted to acquire operability in plants and other eukaryotes.

Figure 13. Avoidance of unwanted self-ligation in palindromic couplings. Figure 13 illustrates that another goal of this invention, in addition to the generation of novel gene pathways, is the subjection of gene pathways – both naturally occurring and man-made – to mutagenesis and selection in order to achieve improved progeny molecules using the instantly disclosed methods of directed evolution (including saturation mutagenesis and synthetic ligation reassembly). In a particular embodiment, as provided by the instant invention, both microbial and plant pathways can be improved by directed evolution, and as shown, the directed evolution process can be performed both on genes prior to linking them into pathways, and on gene pathways themselves.

Figure 14. Conversion of Microbial Pathways to Eukaryotic Pathways. In a particular embodiment, this invention provides that microbial pathways can be converted to pathways operable in plants and other eukaryotic species by the introduction of regulatory sequences that function in those species. Preferred regulatory sequences include promoters, operators, and activator binding sites. As shown, a preferred method of achieving the introduction of such serviceable regulatory sequences is in the form of nucleic acid building blocks, particularly through the use of couplings in ligation reassembly processes. These couplings in Fig. 14 are marked with the letters A, B, C, D and F.

Fig. 15. Engineering of differentially activatable stacked traits in novel transgenic plants using directed evolution and holistic whole cell monitoring. It is a goal of this invention to provide ways of introducing differentially activatable stacked traits into a transgenic cell or organism, the effects of which is holistically monitored. Figure 15 illustrates an approach that may be taken to introduce a plurality of stacked traits into an organism, such as but not limited to a plant, and to carry out holistic whole cell or organism monitoring. Holistic monitoring can include methods pertaining to genomics, RNA profiling, proteomics, metabolomics, and lipid profiling.

Fig. 16. Differential Activation of Selected Traits Can Be Achieved by Adjusting and Controlling the Environment of the Traits. In a particular embodiment,

this invention provides that stacked traits can be introduced into an organism that are differentially activatable, allowing screening under various conditions. Figure 16 illustrates an example in which the stacked traits comprise genetically introduced enzymes. In this example, the enzymes can be selectively and differentially activated by adjusting the environment to which they are exposed.

Fig. 17. Desired or improved traits for harvesting, processing, and storage conditions. One of the goals of this invention is to provide a method that allows the generation of recombinant proteins with desired or improved activities. In a particular embodiment, as illustrated in this figure, a potential application of this method is screening transgenic cells for various responses to harvesting, processing, and storage conditions of biological reagents and strains. The transgenic cells have had stacked traits that are differentially activatable introduced. Screening methods that pertain to methods of genomics, proteomics, RNA profiling, metabolomics, and lipid profiling can be utilized and assessed under various specific conditions that include but are not limited to variations in pH, temperature, and other environmental conditions.

Fig. 18. Mutagenesis and production of a transgenic organism. In another embodiment of this invention, it provides a general method to introduce a library of mutagenized nucleotide sequences (e.g., saturation mutagenesis and/or ligation reassembly) into an organism, and to screen the transgenic organisms for various holistic phenotypes (preferably using a high throughput method). Optionally, mutations can be combined and the organisms rescreened and/or a second library can be introduced into the transgenic organisms and the process repeated. In a preferred embodiment, the starting population is comprised of an organism strain to be subjected to improvement or evolution in order to produce a resultant population comprised of an improved organism strain that has a desired trait.

Fig. 19 Gene Product Processing. Figure 19 illustrates that various processing or decorating steps occur to a gene product prior to it being active. This is a schematic of various processing steps that render a product active or inactive. Once a gene product is

active it can be differentially expressed and in certain cases modifications in its activities or properties can be screened.

Fig. 20. Differential Activation of Selected Precursor (Inactive) Gene Products. Figure 20 is a schematic that illustrates post-translational modifications as a potential process that differentially activates gene products. Differential activation of gene products should be considered when designing screening assays. In screening assays, a transgenic organism may not be selected if the gene product has been inactivated due to post-translational effects such as proteolytic cleavage.

Fig. 21. Production of an improved organism or strain that has a desired trait.. In another embodiment of this invention, it provides a general method to introduce a library of mutagenized nucleotide sequences into an organism, and to screen the transgenic organisms or strain for various phenotypes (preferably using a high throughput method). Screening methods that pertain to methods of genomics, proteomics, RNA profiling, metabolomics, and lipid profiling can be utilized to identify a subset of desired mutants, such as "up-mutants". Optionally, mutations can be combined and the organisms rescreened and/or a second library can be introduced into the transgenic organisms and the process repeated. In a preferred embodiment, the starting population is comprised of an organism strain to be subjected to improvement or evolution in order to produce a resultant population comprised of an improved organism strain that has a desired trait.

Fig. 22. Reassortment of polynucleotide sequences to produce an improved sequence that has a desired trait. Another goal of this invention is to provide a method to prepare mutagenized polynucleotides, to screen the polynucleotide products, and thereby produce an improved sequence with a desired trait. For example, as illustrated in Figure 22, mutagenized polynucleotides can be generated by *in vivo* based reassortment methods such as transposon-based or homologous recombination-based methods. Subsequently, the transgenic organisms can be screened to select a desirable subset of mutants (such as those with an enhanced trait or "up mutant"). The subset of organisms can be selected and

various mutations can be combined. The resultant strain can undergo further rounds of selection for an "up mutant" and/or the improved genomic sequence can be selected and determined.

Fig. 23. **Strain Improvement.** Figure 23 further illustrates the utility of this invention for the generation of improved strains or organisms. This schematic illustratively compares classical and modified classical genetic methods with a method provided in this invention. This invention provides for the generation of strains that harbor more mutations than are typically harbored by strains generated by classical genetic approaches. The generation of strains with numerous mutations and subsequent screening of such strains will allow for the selection of improved strains. As illustrated in this figure, an embodiment of this invention is to generate random clones (e.g., that are a result of three levels of mutagenesis), create transgenic organisms upon the transfer of these clones in a high throughput process, allow in vivo recombination due to homologous recombination, transposon insertion, or suicide plasmids, and identify strains with improved characteristics by screening. Subsequently, the clones that rendered improved characteristics could be identified and combined into one strain with the goal of generating an improved strain due to multiple genetic mutations.

Fig. 24. **Iterative Strain Improvement.** This figure illustrates how this invention provides a method for iterative strain improvement by allowing multiple rounds of mutagenesis, recombination, and selection. In this schematic, a library from an organism is subjected to mutagenesis and then transformed into a parent organism. Once in the cell, additional variation is introduced by *in vivo* recombination (e.g., homologous recombination). Resultant strains are screened for a desired or enhanced trait (an "up mutant") and the mutations are identified and sequenced. Subsequently, various set or subsets of identified clones can be recombined to create further strain improvements.

Fig. 25. **Illustrative diagram for the introduction of mutations for genome site saturated mutagenesis.** In one sense, this method permits the targeted construction of markerless deletions, insertions, and point mutations into a genome (such as a bacterial chromosome) for genome site saturation mutagenesis. Libraries of genomes can be mutagenized (and multiply mutagenized) and introduced into cells, allowing

recombination with genomic alleles. For example as illustrated in this diagram, a suicide plasmid that carries a mutant allele and the recognition site of the yeast meganuclease I-SceI, can be inserted into a genome by homologous recombination between the mutant and the wild-type alleles. Further recombination results in either a mutant or a wildtype chromosome. Pools of mutants generated from the same genome fragment can be combined and stored in one position of an array such that every fragment of the genome can be mutated to saturation.

Figure 26. Producing polynucleotides via interrupted synthesis methods. An embodiment of this invention provides for the production of chimeric/mutagenized polynucleotides (including coding and noncoding regions) generated by incomplete extension. Incomplete extension can be used to generate intermediate products of varying length that ultimately may be utilized to generate pools of chimeric/mutagenized polynucleotides. Various methods can be utilized to interrupt synthesis of nucleic acids: abbreviated annealing times (as exemplified in Figure 27), decreased dNTP concentrations, multiple monobinders priming one polybinder template, template chemistry (such as using a template with chemically modified bases), a DNA polymerase with decreased activity, and/or the use of modified nucleotides during synthesis (such as ddCTP).

Figure 27. Utilizing PCR cycles with abbreviated annealing times for interrupted synthesis. An embodiment of this invention provides for the production of chimeric/mutagenized polynucleotides (including coding and noncoding regions) generated by interrupted synthesis methods. Variations of standard PCR cycles that utilize abbreviated annealing times is one method that can lead to incomplete extension. As illustrated, there are numerous possible variations (such as, but not limited to, variations 1 – 5) that could be utilized.

Figure 28. Example of a flow chart that is serviceable for performing computer-aided analysis according to this invention.

E. DEFINITIONS OF TERMS

In order to facilitate understanding of the examples provided herein, certain frequently occurring methods and/or terms will be described.

The term "**agent**" is used herein to denote a chemical compound, a mixture of chemical compounds, an array of spatially localized compounds (e.g., a VLSIPS peptide array, polynucleotide array, and/or combinatorial small molecule array), biological macromolecule, a bacteriophage peptide display library, a bacteriophage antibody (e.g., scFv) display library, a polysome peptide display library, or an extract made form biological materials such as bacteria, plants, fungi, or animal (particular mammalian) cells or tissues. Agents are evaluated for potential activity as anti-neoplastics, anti-inflammatories or apoptosis modulators by inclusion in screening assays described hereinbelow. Agents are evaluated for potential activity as specific protein interaction inhibitors (i.e., an agent which selectively inhibits a binding interaction between two predetermined polypeptides but which doe snot substantially interfere with cell viability) by inclusion in screening assays described hereinbelow.

An "**ambiguous base requirement**" in a restriction site refers to a nucleotide base requirement that is not specified to the fullest extent, i.e. that is not a specific base (such as, in a non-limiting exemplification, a specific base selected from A, C, G, and T), but rather may be any one of at least two or more bases. Commonly accepted abbreviations that are used in the art as well as herein to represent ambiguity in bases include the following: **R** = G or A; **Y** = C or T; **M** = A or C; **K** = G or T; **S** = G or C; **W** = A or T; **H** = A or C or T; **B** = G or T or C; **V** = G or C or A; **D** = G or A or T; **N** = A or C or G or T.

The term "**amino acid**" as used herein refers to any organic compound that contains an amino group (-NH_2) and a carboxyl group (-COOH); preferably either as free groups or alternatively after condensation as part of peptide bonds. The "**twenty naturally encoded polypeptide-forming alpha-amino acids**" are understood in the art and refer to: alanine (ala or A), arginine (arg or R), asparagine (asn or N), aspartic acid (asp or D), cysteine (cys or C), gluatamic acid (glu or E), glutamine (gln or Q), glycine (gly or G), histidine (his or H), isoleucine (ile or I), leucine (leu or L), lysine (lys or K), methionine (met or M), phenylalanine (phe or F), proline (pro or P), serine (ser or S), threonine (thr or T), tryptophan (trp or W), tyrosine (tyr or Y), and valine (val or V).

The term "**amplification**" means that the number of copies of a polynucleotide is increased.

The term "**antibody**", as used herein, refers to intact immunoglobulin molecules, as well as fragments of immunoglobulin molecules, such as Fab, Fab', $(Fab')_2$, Fv, and SCA fragments, that are capable of binding to an epitope of an antigen. These antibody fragments, which retain some ability to selectively bind to an antigen (*e.g.*, a polypeptide antigen) of the antibody from which they are derived, can be made using well known methods in the art (see, e.g., **Harlow and Lane**, *supra*), and are described further, as follows.

(1) An Fab fragment consists of a monovalent antigen-binding fragment of an antibody molecule, and can be produced by digestion of a whole antibody molecule with the enzyme papain, to yield a fragment consisting of an intact light chain and a portion of a heavy chain.

(2) An Fab' fragment of an antibody molecule can be obtained by treating a whole antibody molecule with pepsin, followed by reduction, to yield a molecule

consisting of an intact light chain and a portion of a heavy chain. Two Fab' fragments are obtained per antibody molecule treated in this manner.

(3) An (Fab')$_2$ fragment of an antibody can be obtained by treating a whole antibody molecule with the enzyme pepsin, without subsequent reduction. A (Fab')$_2$ fragment is a dimer of two Fab' fragments, held together by two disulfide bonds.

(4) An Fv fragment is defined as a genetically engineered fragment containing the variable region of a light chain and the variable region of a heavy chain expressed as two chains.

(5) An single chain antibody ("SCA") is a genetically engineered single chain molecule containing the variable region of a light chain and the variable region of a heavy chain, linked by a suitable, flexible polypeptide linker.

The term "**Applied Molecular Evolution**" ("AME") means the application of an evolutionary design algorithm to a specific, useful goal. While many different library formats for AME have been reported for polynucleotides, peptides and proteins (phage, lacI and polysomes), none of these formats have provided for recombination by random cross-overs to deliberately create a combinatorial library.

A molecule that has a "**chimeric property**" is a molecule that is: 1) in part homologous and in part heterologous to a first reference molecule; while 2) at the same time being in part homologous and in part heterologous to a second reference molecule; without 3) precluding the possibility of being at the same time in part homologous and in part heterologous to still one or more additional reference molecules. In a non-limiting embodiment, a chimeric molecule may be prepared by assemblying a reassortment of partial molecular sequences. In a non-limiting aspect, a chimeric polynucleotide molecule may be prepared by synthesizing the chimeric polynucleotide using plurality of molecular templates, such that the resultant chimeric polynucleotide has properties of a plurality of templates.

The term "**cognate**" as used herein refers to a gene sequence that is evolutionarily and functionally related between species. For example, but not limitation, in the human genome the human CD4 gene is the cognate gene to the mouse 3d4 gene, since the sequences and structures of these two genes indicate that they are highly homologous and both genes encode a protein which functions in signaling T cell activation through MHC class II-restricted antigen recognition.

A "**comparison window,**" as used herein, refers to a conceptual segment of at least 20 contiguous nucleotide positions wherein a polynucleotide sequence may be compared to a reference sequence of at least 20 contiguous nucleotides and wherein the portion of the polynucleotide sequence in the comparison window may comprise additions or deletions (i.e., gaps) of 20 percent or less as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. Optimal alignment of sequences for aligning a comparison window may be conducted by the local homology algorithm of Smith (**Smith and Waterman,** *Adv Appl Math*, 1981; **Smith and Waterman,** J Teor Biol, 1981; **Smith and Waterman,** *J Mol Biol*, 1981; **Smith** et al, *J Mol Evol*, 1981), by the homology alignment algorithm of Needleman (**Needleman and Wuncsch**, 1970), by the search of similarity method of Pearson (Pearson and Lipman, 1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package Release 7.0, Genetics Computer Group, 575 Science Dr., Madison, WI), or by inspection, and the best alignment (i.e., resulting in the highest percentage of homology over the comparison window) generated by the various methods is selected.

As used herein, the term "**complementarity-determining region**" and "**CDR**" refer to the art-recognized term as exemplified by the Kabat and Chothia CDR definitions also generally known as supervariable regions or hypervariable loops (**Chothia and Lesk**, 1987; **Clothia** et al, 1989; **Kabat et al**, 1987; and **Tramontano** et al, 1990). Variable region domains typically comprise the amino-terminal approximately 105-115 amino acids of a naturally-occurring immunoglobulin chain (e.g., amino acids 1-110), although variable domains somewhat shorter or longer are also suitable for forming single-chain antibodies.

"**Conservative amino acid substitutions**" refer to the interchangeability of residues having similar side chains. For example, a group of amino acids having aliphatic side chains is glycine, alanine, valine, leucine, and isoleucine; a group of amino acids having aliphatic-hydroxyl side chains is serine and threonine; a group of amino acids having amide-containing side chains is asparagine and glutamine; a group of amino acids having aromatic side chains is phenylalanine, tyrosine, and tryptophan; a group of amino acids having basic side chains is lysine, arginine, and histidine; and a group of amino acids having sulfur-containing side chains is cysteine and methionine. Preferred conservative amino acids substitution groups are : valine-leucine-isoleucine, phenylalanine-tyrosine, lysine-arginine, alanine-valine, and asparagine-glutamine.

The term "**corresponds to**" is used herein to mean that a polynucleotide sequence is homologous (i.e., is identical, not strictly evolutionarily related) to all or a portion of a reference polynucleotide sequence, or that a polypeptide sequence is identical to a reference polypeptide sequence. In contradistinction, the term "complementary to" is used herein to mean that the complementary sequence is homologous to all or a portion of a reference polynucleotide sequence. For illustration, the nucleotide sequence "TATAC" corresponds to a reference "TATAC" and is complementary to a reference sequence "GTATA."

The term "**degrading effective**" amount refers to the amount of enzyme which is required to process at least 50% of the substrate, as compared to substrate not contacted with the enzyme. Preferably, at least 80% of the substrate is degraded.

As used herein, the term "**defined sequence framework**" refers to a set of defined sequences that are selected on a non-random basis, generally on the basis of experimental data or structural data; for example, a defined sequence framework may comprise a set of amino acid sequences that are predicted to form a ß-sheet structure or may comprise a leucine zipper heptad repeat motif, a zinc-finger domain, among other variations. A "defined sequence kernal" is a set of sequences which encompass a limited scope of variability. Whereas (1) a completely random 10-mer sequence of the 20 conventional amino acids can be any of $(20)^{10}$ sequences, and (2) a pseudorandom 10-mer sequence of the 20 conventional amino acids can be any of $(20)^{10}$ sequences but will exhibit a bias for certain residues at certain positions and/or overall, (3) a defined sequence kernal is a subset of sequences if each residue position was allowed to be any of the allowable 20 conventional amino acids (and/or allowable unconventional amino/imino acids). A defined sequence kernal generally comprises variant and invariant residue positions and/or comprises variant residue positions which can comprise a residue selected from a defined subset of amino acid residues), and the like, either segmentally or over the entire length of the individual selected library member sequence. Defined sequence kernels can refer to either amino acid sequences or polynucleotide sequences. Of illustration and not limitation, the sequences $(NNK)_{10}$ and $(NNM)_{10}$, wherein N represents A, T, G, or C; K represents G or T; and M represents A or C, are defined sequence kernels.

"**Digestion**" of DNA refers to catalytic cleavage of the DNA with a restriction enzyme that acts only at certain sequences in the DNA. The various restriction enzymes used herein are commercially available and their reaction conditions, cofactors and other requirements were used as would be known to the ordinarily skilled artisan. For analytical purposes, typically 1 μg of plasmid or DNA fragment is used with about 2 units of enzyme in about 20 μl of buffer solution. For the purpose of isolating DNA fragments for plasmid construction, typically 5 to 50 μg of DNA are digested with 20 to 250 units of enzyme in a larger volume. Appropriate buffers and substrate amounts for particular restriction enzymes are specified by the manufacturer. Incubation times of about 1 hour at 37°C are ordinarily used, but may vary in accordance with the supplier's instructions. After digestion the reaction is electrophoresed directly on a gel to isolate the desired fragment.

"**Directional ligation**" refers to a ligation in which a 5' end and a 3' end of a polynuclotide are different enough to specify a preferred ligation orientation. For example, an otherwise untreated and undigested PCR product that has two blunt ends will typically not have a preferred ligation orientation when ligated into a cloning vector digested to produce blunt ends in its multiple cloning site; thus, directional ligation will typically not be displayed under these circumstances. In contrast, directional ligation will typically displayed when a digested PCR product having a 5' *Eco*R I-treated end and a 3' *Bam*H I-is ligated into a cloning vector that has a multiple cloning site digested with *Eco*R I and *Bam*H I.

The term "**DNA shuffling**" is used herein to indicate recombination between substantially homologous but non-identical sequences, in some embodiments DNA shuffling may involve crossover via non-homologous recombination, such as via cer/lox and/or flp/frt systems and the like.

As used in this invention, the term "**epitope**" refers to an antigenic determinant on an antigen, such as a phytase polypeptide, to which the paratope of an antibody, such as an phytase-specific antibody, binds. Antigenic determinants usually consist of chemically active surface groupings of molecules, such as amino acids or sugar side chains, and can have specific three-dimensional structural characteristics, as well as specific charge characteristics. As used herein "**epitope**" refers to that portion of an antigen or other macromolecule capable of forming a binding interaction that interacts with the variable region binding body of an antibody. Typically, such binding interaction is manifested as an intermolecular contact with one or more amino acid residues of a CDR.

The terms "**fragment**", "derivative" and "analog" when referring to a reference polypeptide comprise a polypeptide which retains at least one biological function or activity that is at least essentially same as that of the reference polypeptide. Furthermore, the terms "fragment", "derivative" or "analog" are exemplified by a "pro-form" molecule, such as a low activity proprotein that can be modified by cleavage to produce a mature enzyme with significantly higher activity.

A method is provided herein for producing from a template polypeptide a set of progeny polypeptides in which a "**full range of single amino acid substitutions**" is represented at each amino acid position. As used herein, "**full range of single amino acid substitutions**" is in reference to the naturally encoded 20 naturally encoded polypeptide-forming alpha-amino acids, as described herein.

The term "**gene**" means the segment of DNA involved in producing a polypeptide chain; it includes regions preceding and following the coding region (leader and trailer) as well as intervening sequences (introns) between individual coding segments (exons).

"Genetic instability", as used herein, refers to the natural tendency of highly repetitive sequences to be lost through a process of reductive events generally involving sequence simplification through the loss of repeated sequences. Deletions tend to involve the loss of one copy of a repeat and everything between the repeats.

The term "**heterologous**" means that one single-stranded nucleic acid sequence is unable to hybridize to another single-stranded nucleic acid sequence or its complement. Thus areas of heterology means that areas of polynucleotides or polynucleotides have areas or regions within their sequence which are unable to hybridize to another nucleic acid or polynucleotide. Such regions or areas are for example areas of mutations.

The term "**homologous**" or "**homeologous**" means that one single-stranded nucleic acid nucleic acid sequence may hybridize to a complementary single-stranded nucleic acid sequence. The degree of hybridization may depend on a number of factors including the amount of identity between the sequences and the hybridization conditions such as temperature and salt concentrations as discussed later. Preferably the region of identity is greater than about 5 bp, more preferably the region of identity is greater than 10 bp.

An immunoglobulin light or heavy chain variable region consists of a "framework" region interrupted by three hypervariable regions, also called CDR's. The extent of the framework region and CDR's have been precisely defined; see "Sequences of Proteins of Immunological Interest" (**Kabat** et al, 1987). The sequences of the framework regions of different light or heavy chains are relatively conserved within a specie. As used herein, a "**human framework region**" is a framework region that is substantially identical (about 85 or more, usually 90-95 or more) to the framework region of a naturally occurring human immunoglobulin. the framework region of an antibody, that is the combined framework regions of the constituent light and heavy chains, serves to position and align the CDR's. The CDR's are primarily responsible for binding to an epitope of an antigen.

The benefits of this invention extend to "**commercial applications**" (or commercial processes), which term is used to include applications in commercial industry proper (or simply industry) as well as non-commercial commercial applications (e.g.

biomedical research at a non-profit institution). Relevant applications include those in areas of diagnosis, medicine, agriculture, manufacturing, and academia.

The term "**identical**" or "**identity**" means that two nucleic acid sequences have the same sequence or a complementary sequence. Thus, "areas of identity" means that regions or areas of a polynucleotide or the overall polynucleotide are identical or complementary to areas of another polynucleotide or the polynucleotide.

The term "**isolated**" means that the material is removed from its original environment (e.g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or enzyme present in a living animal is not isolated, but the same polynucleotide or enzyme, separated from some or all of the coexisting materials in the natural system, is isolated. Such polynucleotides could be part of a vector and/or such polynucleotides or enzymes could be part of a composition, and still be isolated in that such vector or composition is not part of its natural environment.

By "**isolated nucleic acid**" is meant a nucleic acid, e.g., a DNA or RNA molecule, that is not immediately contiguous with the 5' and 3' flanking sequences with which it normally is immediately contiguous when present in the naturally occurring genome of the organism from which it is derived. The term thus describes, for example, a nucleic acid that is incorporated into a vector, such as a plasmid or viral vector; a nucleic acid that is incorporated into the genome of a heterologous cell (or the genome of a homologous cell, but at a site different from that at which it naturally occurs); and a nucleic acid that exists as a separate molecule, e.g., a DNA fragment produced by PCR amplification or restriction enzyme digestion, or an RNA molecule produced by *in vitro* transcription. The term also describes a recombinant nucleic acid that forms part of a hybrid gene encoding additional polypeptide sequences that can be used, for example, in the production of a fusion protein.

As used herein "**ligand**" refers to a molecule, such as a random peptide or variable segment sequence, that is recognized by a particular receptor. As one of skill in the art will recognize, a molecule (or macromolecular complex) can be both a receptor and a ligand. In general, the binding partner having a smaller molecular weight is referred to as the ligand and the binding partner having a greater molecular weight is referred to as a receptor.

"Ligation" refers to the process of forming phosphodiester bonds between two double stranded nucleic acid fragments (**Sambrook** et al, 1982, p. 146; **Sambrook**, 1989). Unless otherwise provided, ligation may be accomplished using known buffers and conditions with 10 units of T4 DNA ligase ("ligase") per 0.5 µg of approximately equimolar amounts of the DNA fragments to be ligated.

As used herein, "**linker**" or "**spacer**" refers to a molecule or group of molecules that connects two molecules, such as a DNA binding protein and a random peptide, and serves to place the two molecules in a preferred configuration, e.g., so that the random peptide can bind to a receptor with minimal steric hindrance from the DNA binding protein.

As used herein, a **"molecular property to be evolved"** includes reference to molecules comprised of a polynucleotide sequence, molecules comprised of a polypeptide sequence, and molecules comprised in part of a polynucleotide sequence and in part of a polypeptide sequence. Particularly relevant - but by no means limiting - examples of molecular properties to be evolved include enzymatic activities at specified conditions, such as related to temperature; salinity; pressure; pH; and concentration of glycerol, DMSO, detergent, &/or any other molecular species with which contact is made in a reaction environment. Additional particularly relevant - but by no means limiting - examples of molecular properties to be evolved include stabilities - e.g. the amount of a residual molecular property that is present after a specified exposure time to a specified environment, such as may be encountered during storage.

The term "**mutations**" includes changes in the sequence of a wild-type or parental nucleic acid sequence or changes in the sequence of a peptide. Such mutations may be point mutations such as transitions or transversions. The mutations may be deletions, insertions or duplications. A mutation can also be a "**chimerization**", which is exemplified in a progeny molecule that is generated to contain part or all of a sequence of one parental molecule as well as part or all of a sequence of at least one other parental molecule. This invention provides for both chimeric polynucleotides and chimeric polypeptides.

As used herein, the degenerate "**N,N,G/T**" nucleotide sequence represents 32 possible triplets, where "N" can be A, C, G or T.

The term "**naturally-occurring**" as used herein as applied to the object refers to the fact that an object can be found in nature. For example, a polypeptide or polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a source in nature and which has not been intentionally modified by man in the laboratory is naturally occurring. Generally, the term naturally occurring refers to an object as present in a non-pathological (un-diseased) individual, such as would be typical for the species.

As used herein, a **"nucleic acid molecule"** is comprised of at least one base or one base pair, depending on whether it is single-stranded or double-stranded, respectively. Furthermore, a nucleic acid molecule may belong exclusively or chimerically to any group of nucleotide-containing molecules, as exemplified by, but not limited to, the following groups of nucleic acid molecules: RNA, DNA, genomic nucleic acids, non-genomic nucleic acids, naturally occurring and not naturally occurring nucleic acids, and synthetic nucleic acids. This includes, by way of non-limiting example, nucleic acids associated with any organelle, such as the mitochondria, ribosomal RNA, and nucleic acid molecules comprised chimerically of one or more components that are not naturally occurring along with naturally occurring components.

Additionally, a **"nucleic acid molecule"** may contain in part one or more non-nucleotide-based components as exemplified by, but not limited to, amino acids and sugars. Thus, by way of example, but not limitation, a ribozyme that is in part nucleotide-based and in part protein-based is considered a **"nucleic acid molecule"**.

In addition, by way of example, but not limitation, a nucleic acid molecule that is labeled with a detectable moiety, such as a radioactive or alternatively a non-radioactive label, is likewise considered a **"nucleic acid molecule"**.

The terms **“nucleic acid sequence coding for”** or a **"DNA coding sequence of"** or a **"nucleotide sequence encoding"** a particular enzyme – as well as other synonymous terms – refer to a DNA sequence which is transcribed and translated into an enzyme when placed under the control of appropriate regulatory sequences. A "promotor sequence" is a DNA regulatory region capable of binding RNA polymerase in a cell and initiating transcription of a downstream (3’ direction) coding sequence. The promoter is part of the DNA sequence. This sequence region has a start codon at its 3’ terminus. The promoter sequence does include the minimum number of bases where elements necessary to initiate transcription at levels detectable above background. However, after the RNA polymerase binds the sequence and transcription is initiated at the start codon (3’ terminus with a promoter), transcription proceeds downstream in the 3’ direction. Within the promotor sequence will be found a transcription initiation site (conveniently defined by mapping with nuclease S1) as well as protein binding domains (consensus sequences) responsible for the binding of RNA polymerase.

The terms **“nucleic acid encoding an enzyme (protein)”** or **“DNA encoding an enzyme (protein)”** or **"polynucleotide encoding an enzyme (protein)"** and other synonymous terms encompasses a polynucleotide which includes only coding sequence for the enzyme as well as a polynucleotide which includes additional coding and/or non-coding sequence.

In one preferred embodiment, a **“specific nucleic acid molecule species”** is defined by its chemical structure, as exemplified by, but not limited to, its primary sequence. In another preferred embodiment, a specific **“nucleic acid molecule species”** is defined by a function of the nucleic acid species or by a function of a product derived from the nucleic acid species. Thus, by way of non-limiting example, a **“specific nucleic acid molecule species”** may be defined by one or more activities or properties attributable to it, including activities or properties attributable its expressed product.

The instant definition of **“assembling a working nucleic acid sample into a nucleic acid library”** includes the process of incorporating a nucleic acid sample into a vector-based collection, such as by ligation into a vector and transformation of a host. A description of relevant vectors, hosts, and other reagents as well as specific non-limiting examples thereof are provided hereinafter. The instant definition of **"assembling a working nucleic acid sample into a nucleic acid library"** also includes the process of incorporating a nucleic acid sample into a non-vector-based collection, such as by ligation to adaptors. Preferably the adaptors can anneal to PCR primers to facilitate amplification by PCR.

Accordingly, in a non-limiting embodiment, a **"nucleic acid library"** is comprised of a vector-based collection of one or more nucleic acid molecules. In another preferred embodiment a **"nucleic acid library"** is comprised of a non-vector-based collection of nucleic acid molecules. In yet another preferred embodiment a **"nucleic acid library"** is comprised of a combined collection of nucleic acid molecules that is in part vector-based and in part non-vector-based. Preferably, the collection of molecules comprising a library is searchable and separable according to individual nucleic acid molecule species.

The present invention provides a **"nucleic acid construct"** or alternatively a **"nucleotide construct"** or alternatively a **"DNA construct"**. The term "construct" is used herein to describe a molecule, such as a polynucleotide (*e.g.*, a phytase polynucleotide) may optionally be chemically bonded to one or more additional molecular moieties, such as a vector, or parts of a vector. In a specific - but by no means limiting - aspect, a nucleotide construct is exemplified by a DNA expression DNA expression constructs suitable for the transformation of a host cell.

An "**oligonucleotide**" (or synonymously an "**oligo**") refers to either a single stranded polydeoxynucleotide or two complementary polydeoxynucleotide strands which may be chemically synthesized. Such synthetic oligonucleotides may or may not have a 5' phosphate. Those that do not will not ligate to another oligonucleotide without adding a phosphate with an ATP in the presence of a kinase. A synthetic oligonucleotide will ligate to a fragment that has not been dephosphorylated. To achieve polymerase-based amplification (such as with PCR), a "**32-fold degenerate oligonucleotide that is comprised of, in series, at least a first homologous sequence, a degenerate N,N,G/T sequence, and a second homologous sequence**" is mentioned. As used in this context, "homologous" is in reference to homology between the oligo and the parental polynucleotide that is subjected to the polymerase-based amplification.

As used herein, the term "**operably linked**" refers to a linkage of polynucleotide elements in a functional relationship. A nucleic acid is "operably linked" when it is placed into a functional relationship with another nucleic acid sequence. For instance, a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the coding sequence. Operably linked means that the DNA sequences being linked are typically contiguous and, where necessary to join two protein coding regions, contiguous and in reading frame.

A coding sequence is "**operably linked to**" another coding sequence when RNA polymerase will transcribe the two coding sequences into a single mRNA, which is then translated into a single polypeptide having amino acids derived from both coding sequences. The coding sequences need not be contiguous to one another so long as the expressed sequences are ultimately processed to produce the desired protein.

As used herein the term "**parental polynucleotide set**" is a set comprised of one or more distinct polynucleotide species. Usually this term fis used in reference to a progeny polynucleotide set which is preferably obtained by mutagenization of the parental set, in which case the terms "**parental**", "**starting**" and "**template**" are used interchangeably.

As used herein the term "**physiological conditions**" refers to temperature, pH, ionic strength, viscosity, and like biochemical parameters which are compatible with a viable organism, and/or which typically exist intracellularly in a viable cultured yeast cell or mammalian cell. For example, the intracellular conditions in a yeast cell grown under typical laboratory culture conditions are physiological conditions. Suitable *in vitro* reaction conditions for *in vitro* transcription cocktails are generally physiological conditions. In general, *in vitro* physiological conditions comprise 50-200 mM NaCl or KCl, pH 6.5-8.5, 20-45 C and 0.001-10 mM divalent cation (e.g., Mg^{++}, Ca^{++}); preferably about 150 mM NaCl or KCl, pH 7.2-7.6, 5 mM divalent cation, and often include 0.01-1.0 percent nonspecific protein (e.g., BSA). A non-ionic detergent (Tween, NP-40, Triton X-100) can often be present, usually at about 0.001 to 2%, typically 0.05-0.2% (v/v).

Particular aqueous conditions may be selected by the practitioner according to conventional methods. For general guidance, the following buffered aqueous conditions may be applicable: 10-250 mM NaCl, 5-50 mM Tris HCl, pH 5-8, with optional addition of divalent cation(s) and/or metal chelators and/or non-ionic detergents and/or membrane fractions and/or anti-foam agents and/or scintillants.

Standard convention (5' to 3') is used herein to describe the sequence of double standed polynucleotides.

The term "**population**" as used herein means a collection of components such as polynucleotides, portions or polynucleotides or proteins. A "mixed population: means a collection of components which belong to the same family of nucleic acids or proteins (i.e., are related) but which differ in their sequence (i.e., are not identical) and hence in their biological activity.

A molecule having a "**pro-form**" refers to a molecule that undergoes any combination of one or more covalent and noncovalent chemical modifications (e.g. glycosylation, proteolytic cleavage, dimerization or oligomerization, temperature-induced or pH-induced conformational change, association with a co-factor, etc.) en route to attain a more mature molecular form having a property difference (e.g. an increase in activity) in comparison with the reference pro-form molecule. When two or more chemical modification (e.g. two proteolytic cleavages, or a proteolytic cleavage and a deglycosylation) can be distinguished en route to the production of a mature molecule, the referemce precursor molecule may be termed a "**pre-pro-form**" molecule.

As used herein, the term "**pseudorandom**" refers to a set of sequences that have limited variability, such that, for example, the degree of residue variability at another position, but any pseudorandom position is allowed some degree of residue variation, however circumscribed.

"**Quasi-repeated units**", as used herein, refers to the repeats to be re-assorted and are by definition not identical. Indeed the method is proposed not only for practically identical encoding units produced by mutagenesis of the identical starting sequence, but also the reassortment of similar or related sequences which may diverge significantly in some regions. Nevertheless, if the sequences contain sufficient homologies to be reassorted by this approach, they can be referred to as "quasi-repeated" units.

As used herein "**random peptide library**" refers to a set of polynucleotide sequences that encodes a set of random peptides, and to the set of random peptides encoded by those polynucleotide sequences, as well as the fusion proteins contain those random peptides.

As used herein, "**random peptide sequence**" refers to an amino acid sequence composed of two or more amino acid monomers and constructed by a stochastic or random process. A random peptide can include framework or scaffolding motifs, which may comprise invariant sequences.

As used herein, "**receptor**" refers to a molecule that has an affinity for a given ligand. Receptors can be naturally occurring or synthetic molecules. Receptors can be employed in an unaltered state or as aggregates with other species. Receptors can be attached, covalently or non-covalently, to a binding member, either directly or via a specific binding substance. Examples of receptors include, but are not limited to, antibodies, including monoclonal antibodies and antisera reactive with specific antigenic determinants (such as on viruses, cells, or other materials), cell membrane receptors, complex carbohydrates and glycoproteins, enzymes, and hormone receptors.

"**Recombinant**" enzymes refer to enzymes produced by recombinant DNA techniques, i.e., produced from cells transformed by an exogenous DNA construct encoding the desired enzyme. "**Synthetic**" enzymes are those prepared by chemical synthesis.

The term "**related polynucleotides**" means that regions or areas of the polynucleotides are identical and regions or areas of the polynucleotides are heterologous.

"Reductive reassortment", as used herein, refers to the increase in molecular diversity that is accrued through deletion (and/or insertion) events that are mediated by repeated sequences.

The following terms are used to describe the sequence relationships between two or more polynucleotides: **"reference sequence,"** **"comparison window,"** **"sequence identity,"** **"percentage of sequence identity,"** and **"substantial identity."**

A **"reference sequence"** is a defined sequence used as a basis for a sequence comparison; a reference sequence may be a subset of a larger sequence, for example, as a segment of a full-length cDNA or gene sequence given in a sequence listing, or may comprise a complete cDNA or gene sequence. Generally, a reference sequence is at least 20 nucleotides in length, frequently at least 25 nucleotides in length, and often at least 50 nucleotides in length. Since two polynucleotides may each (1) comprise a sequence (i.e., a portion of the complete polynucleotide sequence) that is similar between the two polynucleotides and (2) may further comprise a sequence that is divergent between the two polynucleotides, sequence comparisons between two (or more) polynucleotides are typically performed by comparing sequences of the two polynucleotides over a "comparison window" to identify and compare local regions of sequence similarity.

"Repetitive Index (RI)", as used herein, is the average number of copies of the quasi-repeated units contained in the cloning vector.

The term **"restriction site"** refers to a recognition sequence that is necessary for the manifestation of the action of a restriction enzyme, and includes a site of catalytic cleavage. It is appreciated that a site of cleavage may or may not be contained within a portion of a restriction site that comprises a low ambiguity sequence (i.e. a sequence containing the principal determinant of the frequency of occurrence of the restriction site). Thus, in many cases, relevant restriction sites contain only a low ambiguity sequence with an internal cleavage site (e.g. G/AATTC in the EcoR I site) or an immediately adjacent cleavage site (e.g. /CCWGG in the EcoR II site). In other cases, relevant restriction enzymes [e.g. the Eco57 I site or CTGAAG(16/14)] contain a low ambiguity sequence (e.g. the CTGAAG sequence in the Eco57 I site) with an external cleavage site (e.g. in the

N_{16} portion of the Eco57 I site). When an enzyme (e.g. a restriction enzyme) is said to "**cleave**" a polynucleotide, it is understood to mean that the restriction enzyme catalyzes or facilitates a cleavage of a polynucleotide.

In a non-limiting aspect, a "**selectable polynucleotide**" is comprised of a 5' terminal region (or end region), an intermediate region (i.e. an internal or central region), and a 3' terminal region (or end region). As used in this aspect, a 5' terminal region is a region that is located towards a 5' polynucleotide terminus (or a 5' polynucleotide end); thus it is either partially or entirely in a 5' half of a polynucleotide. Likewise, a 3' terminal region is a region that is located towards a 3' polynucleotide terminus (or a 3' polynucleotide end); thus it is either partially or entirely in a 3' half of a polynucleotide. As used in this non-limiting exemplification, there may be sequence overlap between any two regions or even among all three regions.

The term "**sequence identity**" means that two polynucleotide sequences are identical (i.e., on a nucleotide-by-nucleotide basis) over the window of comparison. The term "percentage of sequence identity" is calculated by comparing two optimally aligned sequences over the window of comparison, determining the number of positions at which the identical nucleic acid base (e.g., A, T, C, G, U, or I) occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison (i.e., the window size), and multiplying the result by 100 to yield the percentage of sequence identity. This "substantial identity", as used herein, denotes a characteristic of a polynucleotide sequence, wherein the polynucleotide comprises a sequence having at least 80 percent sequence identity, preferably at least 85 percent identity, often 90 to 95 percent sequence identity, and most commonly at least 99 percent sequence identity as compared to a reference sequence of a comparison window of at least 25-50 nucleotides, wherein the percentage of sequence identity is calculated by comparing the reference sequence to the polynucleotide sequence which may include deletions or additions which total 20 percent or less of the reference sequence over the window of comparison.

As known in the art "**similarity**" between two enzymes is determined by comparing the amino acid sequence and its conserved amino acid substitutes of one enzyme to the sequence of a second enzyme. Similarity may be determined by procedures which are well-known in the art, for example, a BLAST program (Basic Local Alignment Search Tool at the National Center for Biological Information).

As used herein, the term "**single-chain antibody**" refers to a polypeptide comprising a V_H domain and a V_L domain in polypeptide linkage, generally liked via a spacer peptide (e.g., [Gly-Gly-Gly-Gly-Ser]$_x$), and which may comprise additional amino acid sequences at the amino- and/or carboxy- termini. For example, a single-chain antibody may comprise a tether segment for linking to the encoding polynucleotide. As an example, a scFv is a single-chain antibody. Single-chain antibodies are generally proteins consisting of one or more polypeptide segments of at least 10 contiguous amino substantially encoded by genes of the immunoglobulin superfamily (e.g., see **Williams and Barclay**, 1989, pp. 361-368, which is incorporated herein by reference), most frequently encoded by a rodent, non-human primate, avian, porcine bovine, ovine, goat, or human heavy chain or light chain gene sequence. A functional single-chain antibody generally contains a sufficient portion of an immunoglobulin superfamily gene product so as to retain the property of binding to a specific target molecule, typically a receptor or antigen (epitope).

The members of a pair of molecules (*e.g.*, an antibody-antigen pair or a nucleic acid pair) are said to "**specifically bind**" to each other if they bind to each other with greater affinity than to other, non-specific molecules. For example, an antibody raised against an antigen to which it binds more efficiently than to a non-specific protein can be described as specifically binding to the antigen. (Similarly, a nucleic acid probe can be described as specifically binding to a nucleic acid target if it forms a specific duplex with the target by base pairing interactions (see above).)

"**Specific hybridization**" is defined herein as the formation of hybrids between a first polynucleotide and a second polynucleotide (e.g., a polynucleotide having a distinct but substantially identical sequence to the first polynucleotide), wherein substantially unrelated polynucleotide sequences do not form hybrids in the mixture.

The term "**specific polynucleotide**" means a polynucleotide having certain end points and having a certain nucleic acid sequence. Two polynucleotides wherein one polynucleotide has the identical sequence as a portion of the second polynucleotide but different ends comprises two different specific polynucleotides.

"**Stringent hybridization conditions**" means hybridization will occur only if there is at least 90% identity, preferably at least 95% identity and most preferably at least 97% identity between the sequences. *See* **Sambrook** et al, 1989, which is hereby incorporated by reference in its entirety.

Also included in the invention are polypeptides having sequences that are "**substantially identical**" to the sequence of a phytase polypeptide, such as one of SEQ ID 1. A "substantially identical" amino acid sequence is a sequence that differs from a reference sequence only by conservative amino acid substitutions, for example, substitutions of one amino acid for another of the same class (*e.g.*, substitution of one hydrophobic amino acid, such as isoleucine, valine, leucine, or methionine, for another, or substitution of one polar amino acid for another, such as substitution of arginine for lysine, glutamic acid for aspartic acid, or glutamine for asparagine).

Additionally a "**substantially identical**" amino acid sequence is a sequence that differs from a reference sequence or by one or more non-conservative substitutions, deletions, or insertions, particularly when such a substitution occurs at a site that is not the active site the molecule, and provided that the polypeptide essentially retains its behavioural properties. For example, one or more amino acids can be deleted from a phytase polypeptide, resulting in modification of the structure of the polypeptide, without significantly altering its biological activity. For example, amino- or carboxyl-terminal amino acids that are not required for phytase biological activity can be removed. Such modifications can result in the development of smaller active phytase polypeptides.

The present invention provides a "**substantially pure enzyme**". The term "**substantially pure enzyme**" is used herein to describe a molecule, such as a polypeptide (*e.g.*, a phytase polypeptide, or a fragment thereof) that is substantially free of other proteins, lipids, carbohydrates, nucleic acids, and other biological materials with which it is naturally associated. For example, a substantially pure molecule, such as a polypeptide, can be at least 60%, by dry weight, the molecule of interest. The purity of the polypeptides can be determined using standard methods including, e.g., polyacrylamide gel electrophoresis (*e.g.*, SDS-PAGE), column chromatography (*e.g.*, high performance liquid chromatography (HPLC)), and amino-terminal amino acid sequence analysis.

As used herein, "**substantially pure**" means an object species is the predominant species present (i.e., on a molar basis it is more abundant than any other individual macromolecular species in the composition), and preferably substantially purified fraction is a composition wherein the object species comprises at least about 50 percent (on a molar basis) of all macromolecular species present. Generally, a substantially pure composition will comprise more than about 80 to 90 percent of all macromolecular species present in the composition. Most preferably, the object species is purified to essential homogeneity (contaminant species cannot be detected in the composition by conventional detection methods) wherein the composition consists essentially of a single macromolecular species. Solvent species, small molecules (<500 Daltons), and elemental ion species are not considered macromolecular species.

As used herein, the term "**variable segment**" refers to a portion of a nascent peptide which comprises a random, pseudorandom, or defined kernal sequence. A variable segment" refers to a portion of a nascent peptide which comprises a random pseudorandom, or defined kernal sequence. A variable segment can comprise both variant and invariant residue positions, and the degree of residue variation at a variant residue position may be limited: both options are selected at the discretion of the practitioner. Typically, variable segments are about 5 to 20 amino acid residues in length (e.g., 8 to 10), although variable segments may be longer and may comprise antibody portions or receptor proteins, such as an antibody fragment, a nucleic acid binding protein, a receptor protein, and the like.

The term "**wild-type**" means that the polynucleotide does not comprise any mutations. A "wild type" protein means that the protein will be active at a level of activity found in nature and will comprise the amino acid sequence found in nature.

The term "**working**", as in "**working sample**", for example, is simply a sample with which one is working. Likewise, a "**working molecule**", for example is a molecule with which one is working.

1.Screening and Selection

1.1. Overview of screening and selection

Screening is, in general, a two-step process in which one first determines which cells do and do not express a screening marker and then physically separates the cells having the desired property. Screening markers include, for example, luciferase, beta-galactosidase, and green fluorescent protein. Screening can also be done by observing a cell holistically including but not limited to utilizing methods pertaining to genomics, RNA profiling, proteomics, metabolomics, and lipidomics as well as observing such aspects of growth as colony size, halo formation, etc. Additionally, screening for production of a desired compound, such as a therapeutic drug or "designer chemical" can be accomplished by observing binding of cell products to a receptor or ligand, such as on a solid support or on a column. Such screening can additionally be accomplished by binding to antibodies, as in an ELISA. In some instances the screening process is preferably automated so as to allow screening of suitable numbers of colonies or cells. Some examples of automated screening devices include fluorescence activated cell sorting (FACS), especially in conjunction with cells immobilized in agarose (see Powell et. al. Bio/Technology 8:333-337 (1990); Weaver et. al. Methods 2:234- 247 (1991)), automated ELISA assays, scintillation proximity assays (Hart, H.E. et al., Molecular Immunol. 16:265-267 (1979)) and the formation of fluorescent, colored or UV absorbing compounds on agar plates or in microtitre wells (Krawiec, S., Devel. Indust. Microbiology 31:103-114 (1990)).

Selection is a form of screening in which identification and physical separation are achieved simultaneously, for example, by expression of a selectable marker, which, in some genetic circumstances, allows cells expressing the marker to survive while other cells die (or vice versa). Selectable markers can include, for example, drug, toxin resistance, or nutrient synthesis genes. Selection is also done by such techniques as growth on a toxic substrate to select for hosts having the ability to detoxify a substrate, growth on a new nutrient source to select for hosts having the ability to utilize that nutrient source, competitive growth in culture based on ability to utilize a nutrient source, etc.

In particular, uncloned but differentially expressed proteins (e.g., those induced in response to new compounds, such as biodegradable pollutants in the medium) can be screened by differential display (Appleyard et al. Mol. Gen. Gent. 247:338-342 (1995)). Hopwood (Phil Trans R. Soc. Lond B 324:549-562) provides a review of screens for antibiotic production. Omura (Microbio. Rev. 50:259-279 (1986) and Nisbet (Ann Rev. Med. Chem. 21:149-157 (1986)) disclose screens for antimicrobial agents, including supersensitive bacteria, detection of beta-lactamase and D,D- carboxypeptidase inhibition, beta-lactamase induction, chromogenic substrates and monoclonal antibody screens.

Antibiotic targets can also be used as screening targets in high throughput screening. Antifungals are typically screened by inhibition of fungal growth. Pharmacological agents can be identified as enzyme inhibitors using plates containing the enzyme and a chromogenic substrate, or by automated receptor assays. Hydrolytic enzymes (e.g., proteases, amylases) can be screened by including the substrate in an agar plate and scoring for a hydrolytic clear zone or by using a colorimetric indicator (Steele et al. Ann. Rev. Microbiol. 45:89-106 (1991)). This can be coupled with the use of stains to detect the effects of enzyme action (such as congo red to detect the extent of degradation of celluloses and hemicelluloses).

Tagged substrates can also be used. For example, lipases and esterases can be screened using different lengths of fatty acids linked to umbelliferyl. The action of lipases or esterases removes this tag from the fatty acid, resulting in a quenching or enhancement of umbelliferyl fluorescence. These enzymes can be screened in microtiter plates by a robotic device.

1.2. High-throughput cellular screening: utilizing various types of "omics"

Functional genomics seeks to discover gene function once nucleotide sequence information is available. Proteomics (the study of protein properties such as expression, post-translational modifications, interactions, etc.) and metabolomics (analysis of metabolite pools) are fast-emerging fields complementing functional genomics, that provide a global, integrated view of cellular processes. The variety of techniques and methods used in this effort include the use of bioinformatics, gene-array chips, mRNA differential display, disease models, protein discovery and expression, and target validation. The ultimate goal of many of these efforts has been to develop high-throughput screens for genes of unknown function. For review see Greenbaum D. et al. Genome Res, 11(9):1463-8 (2001).

1.2.1 Genomics

An embodiment of this invention provides for cellular screening; in a particular embodiment, cellular screening may include genomics. "High throughput genomics" refers to application of genomic or genetic data or analysis techniques that use microarrays or other genomic technologies to rapidly identify large numbers of genes or proteins, or distinguish their structure, expression or function from normal or abnormal cells or tissues. An observer can be a person viewing a slide with a microscope or an observer who views digital images. Alternatively, an observer can be a computer-based image analysis system, which automatically observes, analyses and quantitates biological arrayed samples with or without user interaction. Genomics can refer to various investigative techniques that are broad in scope but often refers to measuring gene expression for multitudes of genes simultaneously. For a review see Lockhart, D.J. and Winzeler, E.A. 2000. Genomics, gene expression and DNA arrays. Nature, 405(6788):827-36.

1.2.1.1. Biological Chips

1.2.1.1.1. General considerations

In one aspect the present invention provides for the use of arrays of oligonucleotide probes immobilized in microfabricated patterns on silica chips for analyzing molecular interactions of biological interest. In some assay formats, the oligonucleotide probe is tethered, i.e., by covalent attachment, to a solid support, and arrays of oligonucleotide probes immobilized on solid supports have been used to detect specific nucleic acid sequences in a target nucleic acid. See, e.g., PCT patent publication Nos. WO 89/10977 and 89/11548. Others have proposed the use of large numbers of oligonucleotide probes to provide the complete nucleic acid sequence of a target nucleic acid but failed to provide an enabling method for using arrays of immobilized probes for this purpose. See U.S. Patent Nos. 5,202,231 and 5,002,867 and PCT patent publication No. WO 93/17126. See U.S. Patent No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092, each of which is incorporated herein by reference. Microfabricated arrays of large numbers of oligonucleotide probes, called "DNA chips" offer great promise for a wide variety of applications. New methods and reagents are required to realize this promise, and the present invention helps meet that need.

1.2.1.1.2. General Strategies for utilizing nucleic acid arrays

The invention provides several strategies employing immobilized arrays of probes for comparing a reference sequence of known sequence with a target sequence showing substantial similarity with the reference sequence, but differing in the presence of, e.g., mutations. In a first embodiment, the invention provides a tiling strategy employing an array of immobilized oligonucleotide probes comprising at least two sets of probes. A first probe set comprises a plurality of probes, each probe comprising a segment of at least three nucleotides exactly complementary to a subsequence of the reference sequence, the segment including at least one interrogation position complementary to a corresponding nucleotide in the reference sequence. A second probe set comprises a corresponding probe for each probe in the first probe set, the corresponding probe in the second probe set being identical to a sequence comprising the corresponding probe from the first probe set or a subsequence of at least three nucleotides thereof that includes the at least one interrogation position, except that the at least one interrogation position is occupied by a different nucleotide in each of the two corresponding probes from the first and second probe sets. The probes in the first probe set have at least two interrogation positions corresponding to two contiguous nucleotides in the reference sequence. One interrogation position corresponds to one of the contiguous nucleotides, and the other interrogation position to the other.

In a second embodiment, the invention provides a tiling strategy employing an array comprising four probe sets. A first probe set comprises a plurality of probes, each probe comprising a segment of at least three nucleotides exactly complementary to a subsequence of the reference sequence, the segment including at least one interrogation position complementary to a corresponding nucleotide in the reference sequence. Second, third and fourth probe sets each comprise a corresponding probe for each probe in the first probe set.

The probes in the second, third and fourth probe sets are identical to a sequence comprising the corresponding probe from the first probe set or a subsequence of at least three nucleotides thereof that includes the at least one interrogation position, except that the at least one interrogation position is occupied by a different nucleotide in each of the four corresponding probes from the four probe sets. The first probe set often has at least 100 interrogation positions corresponding to 100 contiguous nucleotides in the reference sequence. Sometimes the first probe set has an interrogation position corresponding to every nucleotide in the reference sequence. The segment of complementarity within the probe set is usually about 9-21 nucleotides. Although probes may contain leading or trailing sequences in addition to the 9-21 sequences, many probes consist exclusively of a 9-21 segment of complementarity.

In a third embodiment, the invention provides immobilized arrays of probes tiled for multiple reference sequences. one such array comprises at least one pair of first and second probe groups, each group comprising first and second sets of probes as defined in the first embodiment. Each probe in the first probe set from the first group is exactly complementary to a subsequence of a first reference sequence, and each probe in the first probe set from the second group is exactly complementary to a subsequence of a second reference sequence.

Thus, the first group of probes are tiled with respect to a first reference sequence and the second group of probes with respect to a second reference sequence. Each group of probes can also include third and fourth sets of probes as defined in the second embodiment. In some arrays of this type, the second reference sequence is a mutated form of the first reference sequence.

In a fourth embodiment, the invention provides arrays for block tiling. Block tiling is a species of the general tiling strategies described above. The usual unit of a block tiling array is a group of probes comprising a wildtype probe, a first set of three mutant probes and a second set of three mutant probes. The wildtype probe comprises a segment of at least three nucleotides exactly complementary to a subsequence of a reference sequence. The segment has at least first and second interrogation positions corresponding to first and second nucleotides in the reference sequence. The probes in the first set of three mutant probes are each identical to a sequence comprising the wildtype probe or a subsequence of at least three nucleotides thereof including the first and second interrogation positions, except in the first interrogation position, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the second set of three mutant probes are each identical to a sequence comprising the wildtype probes or a subsequence of at least three nucleotides thereof including the first and second interrogation positions, except in the second interrogation position, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe.

In a fifth embodiment, the invention provides methods of comparing a target sequence with a reference sequence using arrays of immobilized pooled probes. The arrays employed in these methods represent a further species of the general tiling arrays noted above. In these methods, variants of a reference sequence differing from the reference sequence in at least one nucleotide are identified and each is assigned a designation. An array of pooled probes is provided, with each pool occupying a separate cell of the array. Each pool comprises a probe comprising a segment exactly complementary to each variant sequence assigned a particular designation.

The array is then contacted with a target sequence comprising a variant of the reference sequence. The relative hybridization intensities of the pools in the array to the target sequence are determined. The identity of the target sequence is deduced from the pattern of hybridization intensities. Often, each variant is assigned a designation having at least one digit and at least one value for the digit. In this case, each pool comprises a probe comprising a segment exactly complementary to each variant sequence assigned a particular value in a particular digit. When variants are assigned successive numbers in a numbering system of base m having n digits, n x (m-1) pooled probes are used are used to assign each variant a designation.

In a sixth embodiment, the invention provides a pooled probe for trellis tiling, a further species of the general tiling strategy. In trellis tiling, the identity of a nucleotide in a target sequence is determined from a comparison of hybridization intensities of three pooled trellis probes. A pooled trellis probe comprises a segment exactly complementary to a subsequence of a reference sequence except at a first interrogation position occupied by a pooled nucleotide N, a second interrogation position occupied by a pooled nucleotide selected from the group of three consisting of (1) M or K, (2) R or Y and (3) S or W, and a third interrogation position occupied by a second pooled nucleotide selected from the group. The pooled nucleotide occupying the second interrogation position comprises a nucleotide complementary to a corresponding nucleotide from the reference sequence when the second pooled probe and reference sequence are maximally aligned, and the pooled nucleotide occupying the third interrogation position comprises a nucleotide complementary to a corresponding nucleotide from the reference sequence when the third pooled probe and the reference sequence are maximally aligned. Standard IUPAC nomenclature is used for describing pooled nucleotides.

In trellis tiling, an array comprises at least first, second and third cells, respectively occupied by first, second and third pooled probes, each according to the generic description above. However, the segment of complementarity, location of interrogation positions, and selection of pooled nucleotide at each interrogation position may or may not differ between the three pooled probes subject to the following constraint. One of the three interrogation positions in each of the three pooled probes must align with the same corresponding nucleotide in the reference sequence.

This interrogation position must be occupied by a N in one of the pooled probes, and a different pooled nucleotide in each of the other two pooled probes.

In a seventh embodiment, the invention provides arrays for bridge tiling. Bridge tiling is a species of the general tiling strategies noted above, in which probes from the first probe set contain more than one segment of complementarity.

In bridge tiling, a nucleotide in a reference sequence is usually determined from a comparison of four probes. A first probe comprises at least first and second segments, each of at least three nucleotides and each exactly complementary to first and second subsequences of a reference sequences. The segments including at least one interrogation position corresponding to a nucleotide in the reference sequence.

Either (1) the first and second subsequences are noncontiguous in the reference sequence, or (2) the first and second subsequences are contiguous and the first and second segments are inverted relative to the first and second subsequences.

The arrays further comprises second, third and fourth probes, which are identical to a sequence comprising the first probe or a subsequence thereof comprising at least three nucleotides from each of the first and second segments, except in the at least one interrogation position, which differs in each of the probes. In a species of bridge tiling, referred to as deletion tiling, the first and second subsequences are separated by one or two nucleotides in the reference sequence.

In an eighth embodiment, the invention provides arrays of probes for multiplex tiling. Multiplex tiling is a strategy, in which the identity of two nucleotides in a target sequence is determined from a comparison of the hybridization intensities of four probes, each having two interrogation positions. Each of the probes comprising a segment of at least 7 nucleotides that is exactly complementary to a subsequence from a reference sequence, except that the segment may or may not be exactly complementary at two interrogation positions. The nucleotides occupying the interrogation positions are selected by the following rules: (1) the first interrogation position is occupied by a different nucleotide in each of the four probes, (2) the second interrogation position is occupied by a different nucleotide in each of the four probes, (3) in first and second probes, the segment is exactly complementary to the subsequence, except at no more than one of the interrogation positions, (4) in third and fourth probes, the segment is exactly complementary to the subsequence, except at both of the interrogation positions.

In a ninth embodiment, the invention provides arrays of immobilized probes including helper mutations. Helper mutations are useful for, e.g., preventing self-annealing of probes having inverted repeats. In this strategy, the identity of a nucleotide in a target sequence is usually determined from a comparison of four probes. A first probe comprises a segment of at least 7 nucleotides exactly complementary to a subsequence of a reference sequence except at one or two positions, the segment including an interrogation position not at the one or two positions. The one or two positions are occupied by helper mutations.

Second, third and fourth mutant probes are each identical to a sequence comprising the wildtype probe or a subsequence thereof including the interrogation position and the one or two positions, except in the interrogation position, which is occupied by a different nucleotide in each of the four probes.

In a tenth embodiment, the invention provides arrays of probes comprising at least two probe sets, but lacking a probe set comprising probes that are perfectly matched to a reference sequence. Such arrays are usually employed in methods in which both reference and target sequence are hybridized to the array. The first probe set comprising a plurality of probes, each probe comprising a segment exactly complementary to a subsequence of at least 3 nucleotides of a reference sequence except at an interrogation position. The second probe set comprises a corresponding probe for each probe in the first probe set, the corresponding probe in the second probe set being identical to a sequence comprising the corresponding probe from the first probe set or a subsequence of at least three nucleotides thereof that includes the interrogation position, except that the interrogation position is occupied by a different nucleotide in each of the two corresponding probes and the complement to the reference sequence.

In an eleventh embodiment, the invention provides methods of comparing a target sequence with a reference sequence comprising a predetermined sequence of nucleotides using any of the arrays described above. The methods comprise hybridizing the target nucleic acid to an array and determining which probes, relative to one another, in the array bind specifically to the target nucleic acid. The relative specific binding of the probes indicates whether the target sequence is the same or different from the reference sequence. In some such methods, the target sequence has a substituted nucleotide relative to the reference sequence in at least one undetermined position, and the relative specific binding of the probes indicates the location of the position and the nucleotide occupying the position in the target sequence. In some methods, a second target nucleic acid is also hybridized to the array. The relative specific binding of the probes then indicates both whether the target sequence is the same or different from the reference sequence, and whether the second target sequence is the same or different from the reference sequence. In some methods, when the array comprises two groups of probes tiled for first and second reference sequences, respectively, the relative specific binding of probes in the first group indicates whether the target sequence is the same or different from the first reference sequence. The relative specific binding of probes in the second group indicates whether the target sequence is the same or different from the second reference sequence. Such methods are particularly useful for analyzing heterologous alleles of a gene. Some methods entail hybridizing both a reference sequence and a target sequence to any of the arrays of probes described above. Comparison of the relative specific binding of the probes to the reference and target sequences indicates whether the target sequence is the same or different from the reference sequence.

In a twelfth embodiment, the invention provides arrays of immobilized probes in which the probes are designed to tile a reference sequence from a human immunodeficiency virus.

Reference sequences from either the reverse transcriptase gene or protease gene of HIV are of particular interest. Some chips further comprise arrays of probes tiling a reference sequence from a 16S RNA or DNA encoding the 16S RNA from a pathogenic microorganism. The invention further provides methods of using such arrays in analyzing a HIV target sequence. The methods are particularly useful where the target sequence has a substituted nucleotide relative to the reference sequence in at least one position, the substitution conferring resistance to a drug use in treating a patient infected with a HIV virus. The methods reveal the existence of the substituted nucleotide. The methods are also particularly useful for analyzing a mixture of undetermined proportions of first and second target sequences from different HIV variants. The relative specific binding of probes indicates the proportions of the first and second target sequences.

In a thirteenth embodiment, the invention provides arrays of probes tiled based on reference sequence from a CFTR gene. A preferred array comprises at least a group of probes comprising a wildtype probe, and five sets of three mutant probes. The wildtype probe is exactly complementary to a subsequence of a reference sequence from a cystic fibrosis gene, the segment having at least five interrogation positions corresponding to five contiguous nucleotides in the reference sequence. The probes in the first set of three mutant probes are each identical to the wildtype probe, except in a first of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the second set of three mutant probes are each identical to the wildtype probe, except in a second of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the third set of three mutant probes are each identical to the wildtype probe, except in a third of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the fourth set of three mutant probes are each identical to the wildtype probe, except in a fourth of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the fifth set of three mutant probes are each identical to the wildtype probe, except in a fifth of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. Preferably, a chip comprises two such groups of probes. The first group comprises a wildtype probe exactly complementary to a first reference sequence, and the second group comprises a wildtype probe exactly complementary to a second reference sequence that is a mutated form of the first reference sequence.

The invention further provides methods of using the arrays of the invention for analyzing target sequences from a CFTR gene. The methods are capable of simultaneously analyzing first and second target sequences representing heterozygous alleles of a CFTR gene.

In a fourteenth embodiment, the invention provides arrays of probes tiling a reference sequence from a p53 gene, an hMLHl gene and/or an MSH2 gene. The invention further provides methods of using the arrays described above to analyze these genes. The method are useful, e.g., for diagnosing patients susceptible to developing cancer.

In a fifteenth embodiment, the invention provides arrays of probes tiling a reference sequence from a mitochondrial genome. The reference sequence may comprise part or all of the D-loop region, or all, or substantially all, of the mitochondrial genome. The invention further provides method of using the arrays described above to analyze target sequences from a mitochondrial genome. The methods are useful for identifying mutations associated with disease, and for forensic, epidemiological and evolutionary studies.

1.2.1.1.3. Specific Strategies for utilizing nucleic acid arrays

The invention provides a number of strategies for comparing a polynucleotide of known sequence (a reference sequence) with variants of that sequence (target sequences).

The comparison can be performed at the level of entire genomes, chromosomes, genes, exons or introns, or can focus on individual mutant sites and immediately adjacent bases. The strategies allow detection of variations, such as mutations or polymorphisms, in the target sequence irrespective whether a particular variant has previously been characterized. The strategies both define the nature of a variant and identify its location in a target sequence.

The strategies employ arrays of oligonucleotide probes immobilized to a solid support. Target sequences are analyzed by determining the extent of hybridization at particular probes in the array. The strategy in selection of probes facilitates distinction between perfectly matched probes and probes showing single-base or other degrees of mismatches.

The strategy usually entails sampling each nucleotide of interest in a target sequence several times, thereby achieving a high degree of confidence in its identity. This level of confidence is further increased by sampling of adjacent nucleotides in the target sequence to nucleotides of interest.

The number of probes on the chip can be quite large (e.g., 10^5-10^6). However, usually only a small proportion of the total number of probes of a given length are represented.

Some advantage of the use of only a small proportion of all possible probes of a given length include: (i) each position in the array is highly informative, whether or not hybridization occurs; (ii) nonspecific hybridization is minimized; (iii) it is straightforward to correlate hybridization differences with sequence differences, particularly with reference to the hybridization pattern of a known standard; and (iv) the ability to address each probe independently during synthesis, using high resolution photolithography, allows the array to be designed and optimized for any sequence. For example the length of any probe can be varied independently of the others.

The present tiling strategies result in sequencing and comparison methods suitable for routine large-scale practice with a high degree of confidence in the sequence output.

1.2.1.1.4. General Tiling Strategies

1.2.1.1.4.1. Selection of Reference Sequence

The chips are designed to contain probes exhibiting complementarity to one or more selected reference sequence whose sequence is known. The chips are used to read a target sequence comprising either the reference sequence itself or variants of that sequence. Target sequences may differ from the reference sequence at one or more positions but show a high overall degree of sequence identity with the reference sequence (e.g., at least 75, 90, 95, 99, 99.9 or 99-99%). Any polynucleotide of known sequence can be selected as a reference sequence. Reference sequences of interest include sequences known to include mutations or polymorphisms associated with phenotypic changes having clinical significance in human patients. For example, the CFTR gene and P53 gene in humans have been identified as the location of several mutations resulting in cystic fibrosis or cancer respectively. Other reference sequences of interest include those that serve to identify pathogenic microorganisms and/or are the site of mutations by which such microorganisms acquire drug resistance (e.g., the HIV reverse transcriptase gene). Other reference sequences of interest include regions where polymorphic variations are known to occur (e.g., the D-loop region of mitochondrial DNA). These reference sequences have utility for, e.g., forensic or epidemiological studies. Other reference sequences of interest include p34 (related to p53), p65 (implicated in breast, prostate and liver cancer), and DNA segments encoding cytochromes P450 (see Meyer et al., Pharmac. Ther. 46, 349-355 (1990)). Other reference sequences of interest include those from the genome of pathogenic viruses (e.g., hepatitis J, B, or Q, herpes virus (e.g., VZV, HSV-1, HAV-6, HSV-II, and CMV, Epstein Barr virus), adenovirus, influenza virus, flaviviruses, echovirus, rhinovirus, coxsackie virus, cornovirus, respiratory syncytial virus, mumps virus, rotavirus, measles virus, rubella virus, parvovirus, vaccinia virus, HTLV virus, dengue virus, papillomavirus, molluscum virus, poliovirus, rabies virus, JC virus and arboviral encephalitis virus. Other reference sequences of interest are from genomes or episomes of pathogenic bacteria, particularly regions that confer drug resistance or allow phylogenic characterization of the host (e.g., 16S rRNA or corresponding DNA). For example, such bacteria include chlanydia, rickettsial bacteria, mycobacteria, staphylococci, treptocci, pneumonococci, meningococci and conococci, klebsiella, proteus, serratia, pseudomonas, legionella, diphtheria, salmonella, bacilli, cholera, tetanus, botulism, anthrax, plague, leptospirosis, and Lymes disease bacteria. Other reference sequences of interest include those in which mutations result in the following autosomal recessive disorders: sickle cell anemia, beta-thalassemia, phenylketonuria, galactosemia, Wilson's disease, hemochromatosis, severe combined immunodeficiency, alpha-l-antitrypsin deficiency, albinism, alkaptonuria, lysosomal storage diseases and Ehlers-Danlos syndrome. Other reference sequences of interest include those in which mutations result in X-linked recessive disorders: hemophilia, glucose-6-phosphate dehydrogenase, agammaglobulimenia, diabetes insipidus, Lesch-Nyhan syndrome, muscular dystrophy, Wiskott-Aldrich syndrome, Fabry's disease and fragile X- syndrome. Other reference sequences of interest includes those in which mutations result in the following autosomal dominant disorders: familial hypercholesterolemia, polycystic kidney disease, Huntingdon's disease, hereditary spherocytosis, Marfan's syndrome, von Willebrand's disease, neurofibromatosis, tuberous sclerosis, hereditary hemorrhagic telangiectasia, familial colonic polyposis, Ehlers-Danlos syndrome, myotonic dystrophy, muscular dystrophy, osteogenesis imperfecta, acute intermittent porphyria, and von Hippel- Lindau disease.

The length of a reference sequence can vary widely from a full-length genome, to an individual chromosome, episome, gene, component of a gene, such as an exon, intron or regulatory sequences, to a few nucleotides. A reference sequence of between about 2, 5, 10, 20, 50, 100, 5000, 1000, 5,000 or 10,000, 20,000 or 100,000 nucleotides is common.

Sometimes only particular regions of a sequence (e.g., exons of a gene) are of interest. In such situations, the particular regions can be considered as separate reference sequences or can be considered as components of a single reference sequence, as matter of arbitrary choice.

A reference sequence can be any naturally occurring, mutant, consensus or purely hypothetical sequence of nucleotides, RNA or DNA. For example, sequences can be obtained from computer data bases, publications or can be determined or conceived de novo. Usually, a reference sequence is selected to show a high degree of sequence identity to envisaged target sequences. Often, particularly, where a significant degree of divergence is anticipated between target sequences, more than one reference sequence is selected. Combinations of wildtype and mutant reference sequences are employed in several applications of the tiling strategy.

1.2.1.1.5. Chip Design

1.2.1.1.5.1. Basic Tiling Strategy

The basic tiling strategy provides an array of immobilized probes for analysis of target sequences showing a high degree of sequence identity to one or more selected reference sequences. The strategy is first illustrated for an array that is subdivided into four probe sets, although it will be apparent that in some situations, satisfactory results are obtained from only two probe sets. A first probe set comprises a plurality of probes exhibiting perfect complementarity with a selected reference sequence. The perfect complementarity usually exists throughout the length of the probe. However, probes having a segment or segments of perfect complementarity that is/are flanked by leading or trailing sequences lacking complementarity to the reference sequence can also be used. Within a segment of complementarity, each probe in the first probe set has at least one interrogation position that corresponds to a nucleotide in the reference sequence. That is, the interrogation position is aligned with the corresponding nucleotide in the reference sequence, when the probe and reference sequence are aligned to maximize complementarity between the two. If a probe has more than one interrogation position, each corresponds with a respective nucleotide in the reference sequence. The identity of an interrogation position and corresponding nucleotide in a particular probe in the first probe set cannot be determined simply by inspection of the probe in the first set. As will become apparent, an interrogation position and corresponding nucleotide is defined by the comparative structures of probes in the first probe set and corresponding probes from additional probe sets.

In principle, a probe could have an interrogation position at each position in the segment complementary to the reference sequence. Sometimes, interrogation positions provide more accurate data when located away from the ends of a segment of complementarity. Thus, typically a probe having a segment of complementarity of length x does not contain more than x-2 interrogation positions. Since probes are typically 9-21 nucleotides, and usually all of a probe is complementary, a probe typically has 1-19 interrogation positions. Often the probes contain a single interrogation position, at or near the center of probe.

For each probe in the first set, there are, for purposes of the present illustration, three corresponding probes from three additional probe sets. Thus, there are four probes corresponding to each nucleotide of interest in the reference sequence. Each of the four corresponding probes has an interrogation position aligned with that nucleotide of interest. Usually, the probes from the three additional probe sets are identical to the corresponding probe from the first probe set with one exception. The exception is that at least one (and often only one) interrogation position, which occurs in the same position in each of the four corresponding probes from the four probe sets, is occupied by a different nucleotide in the four probe sets. For example, for an A nucleotide in the reference sequence, the corresponding probe from the first probe set has its interrogation position occupied by a T, and the corresponding probes from the additional three probe sets have their respective interrogation positions occupied by A, C, or G, a different nucleotide in each probe. Of course, if a probe from the first probe set comprises trailing or flanking sequences lacking complementarity to the reference sequences, these sequences need not be present in corresponding probes from the three additional sets. Likewise corresponding probes from the three additional sets can contain leading or trailing sequences outside the segment of complementarity that are not present in the corresponding probe from the first probe set. Occasionally, the probes from the additional three probe set are identical (with the exception of interrogation position(s)) to a contiguous subsequence of the full complementary segment of the corresponding probe from the first probe set. In this case, the subsequence includes the interrogation position and usually differs from the full-length probe only in the omission of one or both terminal nucleotides from the termini of a segment of complementarity.

That is, if a probe from the first probe set has a segment of complementarity of length n, corresponding probes from the other sets will usually include a subsequence of the segment of at least length n-2. Thus, the subsequence is usually at least 3, 4, 7, 9, 15, 21, or 25 nucleotides long, most typically, in the range of 9-21 nucleotides. The subsequence should be sufficiently long to allow a probe to hybridize detectably more strongly to a variant of the reference sequence mutated at the interrogation position than to the reference sequence.

The probes can be oligodeoxyribonucleotides or oligoribonucleotides, or any modified forms of these polymers that are capable of hybridizing with a target nucleic sequence by complementary base-pairing. Complementary base pairing means sequence-specific base pairing which includes e.g., Watson-Crick base pairing as well as other forms of base pairing such as Hoogsteen base pairing. Modified forms include 2□-0-methyl oligoribonucleotides and so-called PNAs, in which oligodeoxyribonucleotides are linked via peptide bonds rather than phophodiester bonds. The probes can be attached by any linkage to a support (e.g., 3□, 5□ or via the base). 3□ attachment is more usual as this orientation is compatible with the preferred chemistry for solid phase synthesis of oligonucleotides.

The number of probes in the first probe set (and as a consequence the number of probes in additional probe sets) depends on the length of the reference sequence, the number of nucleotides of interest in the reference sequence and the number of interrogation positions per probe. In general, each nucleotide of interest in the reference sequence requires the same interrogation position in the four sets of probes.

Consider, as an example, a reference sequence of 100 nucleotides, 50 of which are of interest, and probes each having a single interrogation position. In this situation, the first probe set requires fifty probes, each having one interrogation position corresponding to a nucleotide of interest in the reference sequence. The second, third and fourth probe sets each have a corresponding probe for each probe in the first probe set, and so each also contains a total of fifty probes. The identity of each nucleotide of interest in the reference sequence is determined by comparing the relative hybridization signals at four probes having interrogation positions corresponding to that nucleotide from the four probe sets.

In some reference sequences, every nucleotide is of interest. In other reference sequences, only certain portions in which variants (e.g., mutations or polymorphisms) are concentrated are of interest. In other reference sequences, only particular mutations or polymorphisms and immediately adjacent nucleotides are of interest. Usually, the first probe set has interrogation positions selected to correspond to at least a nucleotide (e.g., representing a point mutation) and one immediately adjacent nucleotide. Usually, the probes in the first set have interrogation positions corresponding to at least 3, 10, 50, 100, 1000, or 20,000 contiguous nucleotides. The probes usually have interrogation positions corresponding to at least 5, 10, 30, 50, 75, 90, 99 or sometimes 100% of the nucleotides in a reference sequence.

Frequently, the probes in the first probe set completely span the reference sequence and overlap with one another relative to the reference sequence. For example, in one common arrangement each probe in the first probe set differs from another probe in that set by the omission of a 3□ base complementary to the reference sequence and the acquisition of a 5□ base complementary to the reference sequence.

For conceptual simplicity, the probes in a set are usually arranged in order of the sequence in a lane across the chip. A lane contains a series of overlapping probes, which represent or tile across, the selected reference sequence. The components of the four sets of probes are usually laid down in four parallel lanes, collectively constituting a row in the horizontal direction and a series of 4-member columns in the vertical direction. Corresponding probes from the four probe sets (i.e., complementary to the same subsequence of the reference sequence) occupy a column.

Each probe in a lane usually differs from its predecessor in the lane by the omission of a base at one end and the inclusion of additional base at the other end. However, this orderly progression of probes can be interrupted by the inclusion of control probes or omission of probes in certain columns of the array. Such columns serve as controls to orient the chip, or gauge the background, which can include target sequence nonspecifically bound to the chip.

The probes sets are usually laid down in lanes such that all probes having an interrogation position occupied by an A form an-A-lane, all probes having an interrogation position occupied by a C form a C-lane, all probes having an interrogation position occupied by a G form a G-lane, and all probes having an interrogation position occupied by a T (or U) form a T lane (or a U lane). Note that in this arrangement there is not a unique correspondence between probe sets and lanes. Thus, the probe from the first probe set is laid down in the A-lane, C-lane, A-lane, A-lane and T-lane for the five columns. The interrogation position on a column of probes corresponds to the position in the target sequence whose identity is determined from analysis of hybridization to the probes in that column. The interrogation position can be anywhere in a probe but is usually at or near the central position of the probe to maximize differential hybridization signals between a perfect match and a single-base mismatch.

For example, for an 11 mer probe, the central position is the sixth nucleotide.

Although the array of probes is usually laid down in rows and columns as described above, such a physical arrangement of probes on the chip is not essential. Provided that the spatial location of each probe in an array is known, the data from the probes can be collected apd processed to yield the sequence of a target irrespective of the physical arrangement of the probes on a chip. In processing the data, the hybridization signals from the respective probes can be reassorted into any conceptual array desired for subsequent data reduction whatever the physical arrangement of probes on the chip.

A range of lengths of probes can be employed in the chips. As noted above, a probe may consist exclusively of a complementary segments, or may have one or more complementary segments juxtaposed by flanking, trailing and/or intervening segments. In the latter situation, the total length of complementary segment(s) is more important than the length of the probe. In functional terms, the complementarity segment(s) of the first probe sets should be sufficiently long to allow the probe to hybridize detectably more strongly to a reference sequence compared with a variant of the reference including a single base mutation at the nucleotide corresponding to the interrogation position of the probe.

Similarly, the complementarity segment(s) in corresponding probes from additional probe sets should be sufficiently long to allow a probe to hybridize detectably more strongly to a variant of the reference sequence having a single nucleotide substitution at the interrogation position relative to the reference sequence. A probe usually has a single complementary segment having a length of at least 3 nucleotides, and more usually at least 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 or bases exhibiting perfect complementarity (other than possibly at the interrogation position(s) depending on the probe set) to the reference sequence. In bridging strategies, where more than one segment of complementarity is present, each segment provides at least three complementary nucleotides to the reference sequence and the combined segments provide at least two segments of three or a total of six complementary nucleotides. As in the other strategies, the combined length of complementary segments is typically from 6-30 nucleotides, and preferably from about 9-21 nucleotides. The two segments are often approximately the same length. Often, the probes (or segment of complementarity within probes) have an odd number of bases, so that an interrogation position can occur in the exact center of the probe.

In some chips, all probes are the same length. Other chips employ different groups of probe sets, in which case the probes are of the same size within a group, but differ between different groups. For example, some chips have one group comprising four sets of probes as described above in which all the probes are 11 mers, together with a second group comprising four sets of probes in which all of the probes are 13 mers. Of course, additional groups of probes can be added.

Thus, some chips contain, e.g., four groups of probes having sizes of 11 mers, 13 mers, 15 mers and 17 mers. Other chips have different size probes within the same group of four probe sets. In these chips, the probes in the first set can vary in length independently of each other. Probes in the other sets are usually the same length as the probe occupying the same column from the first set. However, occasionally different lengths of probes can be included at the same column position in the four lanes. The different length probes are included to equalize hybridization signals from probes irrespective of whether A-T or C-G bonds are formed at the interrogation position.

The length of probe can be important in distinguishing between a perfectly matched probe and probes showing a single- base mismatch with the target sequence. The discrimination is usually greater for short probes. Shorter probes are usually also less susceptible to formation of secondary structures.

However, the absolute amount of target sequence bound, and hence the signal, is greater for larger probes. The probe length representing the optimum compromise between these competing considerations may vary depending on inter alia the GC content of a particular region of the target DNA sequence, secondary structure, synthesis efficiency and cross- hybridization. In some regions of the target, depending on hybridization conditions, short probes (e.g., 11 mers) may provide information that is inaccessible from longer probes (e.g., 19 mers) and vice versa. Maximum sequence information can be read by including several groups of different sized probes on the chip as noted above. However, for many regions of the target sequence, such a strategy provides redundant information in that the same sequence is read multiple times from the different groups of probes. Equivalent information can be obtained from a single group of different sized probes in which the sizes are selected to maximize readable sequence at particular regions of the target sequence. The strategy of customizing probe length within a single group of probe sets minimizes the total number of probes required to read a particular target sequence. This leaves ample capacity for the chip to include probes to other reference sequences.

The invention provides an optimization block which allows systematic variation of probe length and interrogation position to optimize the selection of probes for analyzing a particular nucleotide in a reference sequence. The block comprises alternating columns of probes complementary to the wildtype target and probes complementary to a specific mutation. The interrogation position is varied between columns and probe length is varied down a column.

Hybridization of the chip to the reference sequence or the mutant form of the reference sequence identifies the probe length and interrogation position providing the greatest differential hybridization signal.

The probes are designed to be complementary to either strand of the reference sequence (e.g., coding or non-coding). some chips contain separate groups of probes, one complementary to the coding strand, the other complementary to the noncoding strand. Independent analysis of coding and noncoding strands provides largely redundant information.

However, the regions of ambiguity in reading the coding strand are not always the same as those in reading the noncoding strand. Thus, combination of the information from coding and noncoding strands increases the overall accuracy of sequencing.

Some chips contain additional probes or groups of probes designed to be complementary to a second reference sequence.

The second reference sequence is often a subsequence of the first reference sequence bearing one or more commonly occurring mutations or interstrain variations. The second group of probes is designed by the same principles as described above except that the probes exhibit complementarity to the second reference sequence. The inclusion of a second group is particular useful for analyzing short subsequences of the primary reference sequence in which multiple mutations are expected to occur within a short distance commensurate with the length of the probes (i.e., two or more mutations within 9 to 21 bases). Of course, the same principle can be extended to provide chips containing groups of probes for any number of reference sequences. Alternatively, the chips may contain additional probe(s) that do not form part of a tiled array as noted above, but rather serves as probe(s) for a conventional reverse dot blot. For example, the presence of mutation can be detected from binding of a target sequence to a single oligomeric probe harboring the mutation. Preferably, an additional probe containing the equivalent region of the wildtype sequence is included as a control.

The chips are read by comparing the intensities of labelled target bound to the probes in an array.

Specifically, a comparison is performed between each lane of probes (e.g., A, C, G and T lanes) at each columnar position (physical or conceptual). For a particular columnar position, the lane showing the greatest hybridization signal is called as the nucleotide present at the position in the target sequence corresponding to the interrogation position in the probes. The corresponding position in the target sequence is that aligned with the interrogation position in corresponding probes when the probes and target are aligned to maximize complementarity. Of the four probes in a column, only one can exhibit a perfect match to the target sequence whereas the others usually exhibit at least a one base pair mismatch. The probe exhibiting a perfect match usually produces a substantially greater hybridization signal than the other three probes in the column and is thereby easily identified. However, in some regions of the target sequence, the distinction between a perfect match and a one-base mismatch is less clear. Thus, a call ratio is established to define the ratio of signal from the best hybridizing probes to the second best hybridizing probe that must be exceeded for a particular target position to be read from the probes. A high call ratio ensures that few if any errors are made in calling target nucleotides, but can result in some nucleotides being scored as ambiguous, which could in fact be accurately read.

A lower call ratio results in fewer ambiguous calls, but can result in more erroneous calls. It has been found that at a call ratio of 1.2 virtually all calls are accurate. However, a small but significant number of bases (e.g., up to about %) may have to be scored as ambiguous.

Although small regions of the target sequence can sometimes be ambiguous, these regions usually occur at the same or similar segments in different target sequences. Thus, for precharacterized mutations, it is known in advance whether that mutation is likely to occur within a region of unambiguously determinable sequence.

An array of probes is most useful for analyzing the reference sequence from which the probes were designed and variants of that sequence exhibiting substantial sequence similarity with the reference sequence (e.g., several single- base mutants spaced over the reference sequence). When an array is used to analyze the exact reference sequence from which it was designed, one probe exhibits a perfect match to the reference sequence, and the other three probes in the same column exhibits single-base mismatches. Thus, discrimination between hybridization signals is usually high and accurate sequence is obtained. High accuracy is also obtained when an array is used for analyzing a target sequence comprising a variant of the reference sequence that has a single mutation relative to the reference sequence, or several widely spaced mutations relative to the reference sequence. At different mutant loci, one probe exhibits a perfect match to the target, and the other three probes occupying the same column exhibit single-base mismatches, the difference (with respect to analysis of the reference sequence) being the lane in which the perfect match occurs.

For target sequences showing a high degree of divergence from the reference strain or incorporating several closely spaced mutations from the reference strain, a single group of probes (i.e., designed with respect to a single reference sequence) will not always provide accurate sequence for the highly variant region of this sequence. At some particular columnar positions, it may be that no single probe exhibits perfect complementarity to the target and that any comparison must be based on different degrees of mismatch between the four probes. Such a comparison does not always allow the target nucleotide corresponding to that columnar position to be called. Deletions in target sequences can be detected by loss of signal from probes having interrogation positions encompassed by the deletion. However, signal may also be lost from probes having interrogation positions closely proximal to the deletion resulting in some regions of the target sequence that cannot be read. Target sequence bearing insertions will also exhibit short regions including and proximal to the insertion that usually cannot be read.

The presence of short regions of difficult-to-read target because of closely spaced mutations, insertions or deletion, does not prevent determination of the remaining sequence of the target as different regions of a target sequence are determined independently. Moreover, such ambiguities as might result from analysis of diverse variants with a single group of probes can be avoided by including multiple groups of probe sets on a chip. For example, one group of probes can be designed based on a full-length reference sequence, and the other groups on subsequences of the reference sequence incorporating frequently occurring mutations or strain variations.

A particular advantage of the present sequencing strategy over conventional sequencing methods is the capacity simultaneously to detect and quantify proportions of multiple target sequences. Such capacity is valuable, e.g., for diagnosis of patients who are heterozygous with respect to a gene or who are infected with a virus, such as HIV, which is usually present in several polymorphic forms. Such capacity is also useful in analyzing targets from biopsies of tumor cells and surrounding tissues. The presence of multiple target sequences is detected from the relative signals of the four probes at the array columns corresponding to the target nucleotides at which diversity occurs. The relative signals at the four probes for the mixture under test are compared with the corresponding signals from a homogeneous reference sequence. An increase in a signal from a probe that is mismatched with respect to the reference sequence, and a corresponding decrease in the signal from the probe which is matched with the reference sequence signal the presence of a mutant strain in the mixture. The extent in shift in hybridization signals of the probes is related to the proportion of a target sequence in the mixture. Shifts in relative hybridization signals can be quantitatively related to proportions of reference and mutant sequence by prior calibration of the chip with seeded mixtures of the mutant and reference sequences. By this means, a chip can be used to detect variant or mutant strains constituting as little as 1, 5, 20, or 25 % of a mixture of stains.

Similar principles allow the simultaneous analysis of multiple target sequences even when none is identical to the reference sequence. For example, with a mixture of two target sequences bearing first and second mutations, there would be a variation in the hybridization patterns of probes having interrogation positions corresponding to the first and second mutations relative to the hybridization pattern with the reference sequence. At each position, one of the probes having a mismatched interrogation position relative to the reference sequence would show an increase in hybridization signal, and the probe having a matched interrogation position relative to the reference sequence would show a decrease in hybridization signal. Analysis of the hybridization pattern of the mixture of mutant target sequences, preferably in comparison with the hybridization pattern of the reference sequence, indicates the presence of two mutant target sequences, the position and nature of the mutation in each strain, and the relative proportions of each strain.

In a variation of the above method, the different components in a mixture of target sequences are differentially labelled before being applied to the array. For example, a variety of fluorescent labels emitting at different wavelength are available. The use of differential labels allows independent analysis of different targets bound simultaneously to the array. For example, the methods permit comparison of target sequences obtained from a patient at different stages of a disease.

1.2.1.1.5.2. Omission of Probes

The general strategy outlined above employs four probes to read each nucleotide of interest in a target sequence. One probe (from the first probe set) shows a perfect match to the reference sequence and the other three probes (from the second, third and fourth probe sets) exhibit a mismatch with the reference sequence and a perfect match with a target sequence bearing a mutation at the nucleotide of interest.

The provision of three probes from the second, third and fourth probe sets allows detection of each of the three possible nucleotide substitutions of any nucleotide of interest. However, in some reference sequences or regions of reference sequences, it is known in advance that only certain mutations are likely to occur. Thus, for example, at one site it might be known that an A nucleotide in the reference sequence may exist as a T mutant in some target sequences but is unlikely to exist as a C or G mutant. Accordingly, for analysis of this region of the reference sequence, one might include only the first and second probe sets, the first probe set exhibiting perfect complementarity to the reference sequence, and the second probe set having an interrogation position occupied by an invariant A residue (for detecting the T mutant). In other situations, one might include the first, second and third probes sets (but not the fourth) for detection of a wildtype nucleotide in the reference sequence and two mutant variants thereof in target sequences. In some chips, probes that would detect silent mutations (i.e., not affecting amino acid sequence) are omitted.

In some chips, the probes from the first probe set are omitted corresponding to some or all positions of the reference sequences. Such chips comprise at least two probe sets. The first probe set has a plurality of probes. Each probe comprises a segment exactly complementary to a subsequence of a reference sequence except in at least one interrogation position. A second probe set has a corresponding probe for each probe in the first probe set.

The corresponding probe in the second probe set is identical to a sequence comprising the corresponding probe form the first probe set or a subsequence thereof that includes the at least one (and usually only one) interrogation position except that the at least one interrogation position is occupied by a different nucleotide in each of the two corresponding probes from the first and second probe sets. A third probe set, if present, also comprises a corresponding probe for each probe in the first probe set except at the at least one interrogation position, which differs in the corresponding probes from the three sets. Omission of probes having a segment exhibiting perfect complementarity to the reference sequence results in loss of control information, i.e., the detection of nucleotides in a target sequence that are the same As those in a reference sequence. However, similar information can be obtained by hybridizing a chip lacking probes from the first probe set to both target and reference sequences. The hybridization can be performed sequentially, or concurrently, if the target and reference are differentially labelled. In this situation, the presence of a mutation is detected by a shift in the background hybridization intensity of the reference sequence to a perfectly matched hybridization signal of the target sequence, rather than by a comparison of the hybridization intensities of probes from the first set with corresponding probes from the second, third and fourth sets.

1.2.1.1.5.3. Wildtype Probe Lane

When the chips comprise four probe sets, as discussed supra, and the probe sets are laid down in four lanes, an A-lane, a C-lane, a G-lane and a T or U-lane, the probe having a segment exhibiting perfect complementarity to a reference sequence varies between the four lanes from one column to another. This does not present any significant difficulty in computer analysis of the data from the chip. However, visual inspection of the hybridization pattern of the chip is sometimes facilitated by provision of an extra lane of probes, in which each probe has a segment exhibiting perfect complementarity to the reference sequence. This segment-is identical to a segment from one of the probes in the other four lanes (which lane depending on the column position). The extra lane of probes (designated the wildtype lane) hybridizes to a target sequence at all nucleotide positions except those in which deviations from the reference sequence occurs. The hybridization pattern of the wildtype lane thereby provides a simple visual indication of mutations.

1.2.1.1.5.4. Deletion, Insertion and Multiple-Mutation Probes

Some chips provide an additional probe set specifically designed for analyzing deletion mutations. The additional probe set comprises a probe corresponding to each probe in the first probe set as described above. However, a probe from the additional probe set differs from the corresponding probe in the first probe set in that the nucleotide occupying the interrogation position is deleted in the probe from the additional probe set. Optionally, the probe from the additional probe set bears an additional nucleotide at one of its termini relative to the corresponding probe from the first probe set. The probe from the additional probe set will hybridize more strongly than the corresponding probe from the first probe set to a target sequence having a single base deletion at the nucleotide corresponding to the interrogation position. Additional probe sets are provided in which not only the interrogation position, but also an adjacent nucleotide is detected.

Similarly, other chips provide additional probe sets for analyzing insertions. For example, one additional probe set has a probe corresponding to each probe in the first probe set as described above. However, the probe in the additional probe set has an extra T nucleotide inserted adjacent to the interrogation position. Optionally, the probe has one fewer nucleotide at one of its termini relative to the corresponding probe from the first probe set. The probe from the additional probe set hybridizes more strongly than the corresponding probe from the first probe set to a target sequence having an A nucleotide inserted in a position adjacent to that corresponding to the interrogation position.

Similar additional probe sets are constructed having C, G or T/U nucleotides inserted adjacent to the interrogation position. Usually, four such probe sets, one for each nucleotide, are used in combination.

Other chips provide additional probes (multiple-mutation probes) for analyzing target sequences having multiple closely spaced mutations. A multiple-mutation probe is usually identical to a corresponding probe from the first set as described above, except in the base occupying the interrogation position, and except at one or more additional positions, corresponding to nucleotides in which substitution may occur in the reference sequence. The one or more additional positions in the multiple mutation probe are occupied by nucleotides complementary to the nucleotides occupying corresponding positions in the reference sequence when the possible substitutions have occurred.

1.2.1.1.5.5. Block Tiling

As noted in the discussion of the general tiling strategy, a probe in the first probe set sometimes has more than one interrogation position. In this situation, a probe in the first probe set is sometimes matched with multiple groups of at least one, and usually, three additional probe sets. Three additional probe sets are used to allow detection of the three possible nucleotide substitutions at any one position. If only certain types of substitution are likely to occur (e.g., transitions), only one or two additional probe sets are required (analogous to the use of probes in the basic tiling strategy). To illustrate for the situation where a group comprises three additional probe sets, a first such group comprises second, third and fourth probe sets, each of which has a probe corresponding to each probe in the first probe set. The corresponding probes from the second, third and fourth probes sets differ from the corresponding probe in the first set at a first of the interrogation positions. Thus, the relative hybridization signals from corresponding probes from the first, second, third and fourth probe sets indicate the identity of the nucleotide in a target sequence corresponding to the first interrogation position. A second group of three probe sets (designated fifth, sixth and seventh probe sets), each also have a probe corresponding to each probe in the first probe set. These corresponding probes differ from that in the first probe set at a second interrogation position. The relative hybridization signals from corresponding probes from the first, fifth, sixth, and seventh probe sets indicate the identity of the nucleotide in the target sequence corresponding to the second interrogation position. As noted above, the probes in the first probe set often have seven or more interrogation positions. If there are seven interrogation positions, there are seven groups of three additional probe sets, each group of three probe sets serving to identify the nucleotide corresponding to one of the seven interrogation positions.

Each block of probes allows short regions of a target sequence to be read. For example, for a block of probes having seven interrogation positions, seven nucleotides in the target sequence can be read. Of course, a chip can contain any number of blocks depending on how many nucleotides of the target are of interest. The hybridization signals for each block can be analyzed independently of any other block. The block tiling strategy can also be combined with other tiling strategies, with different parts of the same reference sequence being tiled by different strategies.

The block tiling strategy offers two advantages over the basic strategy in which each probe in the first set has a single interrogation position. One advantage is that the same sequence information can be obtained from fewer probes. A second advantage is that each of the probes constituting a block (i.e., a probe from the first probe set and a corresponding probe from each of the other probe sets) can have identical 3□ and 5□ sequences, with the variation confined to a central segment containing the interrogation positions. The identity of 3□ sequence between different probes simplifies the strategy for solid phase synthesis of the probes on the chip and results in more uniform deposition of the different probes on the chip, thereby in turn increasing the uniformity of signal to noise ratio for different regions of the chip. A third advantage is that greater signal uniformity is achieved within a block.

1.2.1.1.5.6. Multiplex Tiling

In the block tiling strategy discussed above, the identity of a nucleotide in a target or reference sequence is determined by comparison of hybridization patterns of one probe having a segment showing a perfect match with that of other probes (usually three other probes) showing a single base mismatch. In multiplex tiling, the identity of at least two nucleotides in a reference or target sequence is determined by comparison of hybridization signal intensities of four probes, two of which have a segment showing perfect complementarity or a single base mismatch to the reference sequence, and two of which have a segment showing perfect complementarity or a double-base mismatch to a segment. The four probes whose hybridization patterns are to be compared each have a segment that is exactly complementary to a reference sequence except at two interrogation positions, in which the segment may or may not be complementary to the reference sequence. The interrogation positions correspond to the nucleotides in a reference or target sequence which are determined by the comparison of intensities. The nucleotides occupying the interrogation positions in the four probes are selected according to the following rule. The first interrogation position is occupied by a different nucleotide in each of the four probes. The second interrogation position is also occupied by a different nucleotide in each of the four probes. In two of the four probes, designated the first and second probes, the segment is exactly complementary to the reference sequence except at not more than one of the two interrogation positions. In other words, one of the interrogation positions is occupied by a nucleotide that is complementary to the corresponding nuclectide from the reference sequence and the other interrogation position may or may not be so occupied. In the other two of the four probes, designated the third and fourth probes, the segment is exactly complementary to the reference sequence except that both interrogation positions are occupied by nucleotides which are noncomplementary to the respective corresponding nucleotides in the reference sequence.

There are number of ways of satisfying these conditions depending on whether the two nucleotides in the reference sequence corresponding to the two interrogation positions are the same or different. If these two nucleotides are different in the reference sequence (probability 3/4), the conditions are satisfied by each of the two interrogation positions being occupied by the same nucleotide in any given probe. For example, in the first probe, the two interrogation positions would both be A, in the second probe, both would be C, in the third probe, each would be G, and in the fourth probe each would be T or U. If the two nucleotides in the reference sequence corresponding to the two interrogation positions are different, the conditions noted above are satisfied by each of the interrogation positions in any one of the four probes being occupied by complementary nucleotides. For example, in the first probe, the interrogation positions could be occupied by A and T, in the second probe by C and G, in the third probe by G and C and in the four probe, by T and A.

When the four probes are hybridized to a target that is the same as the reference sequence or differs from the reference sequence at one (but not both) of the interrogation positions, two of the four probes show a double-mismatch with the target and two probes show a single mismatch. The identity of probes showing these different degrees of mismatch can be determined from the different hybridization signals.

From the identity of the probes showing the different degrees of mismatch, the nucleotides occupying both of the interrogation positions in the target sequence can be deduced.

For ease of illustration, the multiplex strategy has been initially described for the situation where there are two nucleotides of interest in a reference sequence and only four probes in an array. Of course, the strategy can be extended to analyze any number of nucleotides in a target sequence by using additional probes. In one variation, each pair of interrogation positions is read from a unique group of four probes. In a block variation, different groups of four probes exhibit the same segment of complementarity with the reference sequence, but the interrogation positions move within a block.

The block and standard multiplex tiling variants can of course be used in combination for different regions of a reference sequence. Either or both variants can also be used in combination with any of the other tiling strategies described.

1.2.1.1.5.7. Helper Mutations

Occasionally small regions of a reference sequence give a low hybridization signal as a result of annealing of probes.

The self-annealing reduces the amount of probe effectively available for hybridizing to the target. Although such regions of the target are generally small and the reduction of hybridization signal is usually not so substantial as to obscure the sequence of this region, this concern can be avoided by the use of probes incorporating helper mutations.

The helper mutation(s) serve to break-up regions of internal complementarity within a probe and thereby prevent annealing.

Usually, one or two helper mutations are quite sufficient for this purpose. The inclusion of helper mutations can be beneficial in any of the tiling strategies noted above. In general each probe having a particular interrogation position has the same helper mutation(s). Thus, such probes have a segment in common which shows perfect complementarity with a reference sequence, except that the segment contains at least one helper mutation (the same in each of the probes) and at least one interrogation position (different in all of the probes). For example, in the basic tiling strategy, a probe from the first probe set comprises a segment containing an interrogation position and showing perfect complementarity with a reference sequence except for one or two helper mutations. The corresponding probes from the second, third and fourth probe sets usually comprise the same segment (or sometimes a subsequence thereof including the helper mutation(s) and interrogation position), except that the base occupying the interrogation position varies in each probe.

Usually, the helper mutation tiling strategy is used in conjunction with one of the tiling strategies described above.

The probes containing helper mutations are used to tile regions of a reference sequence otherwise giving low hybridization signal (e.g., because of self-complementarity), and the alternative tiling strategy is used to tile intervening regions.

1.2.1.1.5.8. Pooling Strategies

Pooling strategies also employ arrays of immobilized probes. Probes are immobilized in cells of an array, and the hybridization signal of each cell can be determined independently of any other cell. A particular cell may be occupied by pooled mixture of probes. Although the identity of each probe in the mixture is known, the individual probes in the pool are not separately addressable. Thus, the hybridization signal from a cell is the aggregate of that of the different probes occupying the cell. In general, a cell is scored as hybridizing to a target sequence if at least one probe occupying the cell comprises a segment exhibiting perfect complementarity to the target sequence.

A simple strategy to show the increased power of pooled strategies over a standard tiling is to create three cells each containing a pooled probe having a single pooled position, the pooled position being the same in each of the pooled probes. At the pooled position, there are two possible nucleotides, allowing the pooled probe to hybridize to two target sequences. In tiling terminology, the pooled position of each probe is an interrogation position. As will become apparent, comparison of the hybridization intensities of the pooled probes from the three cells reveals the identity of the nucleotide in the target sequence corresponding to the interrogation position (i.e., that is matched with the interrogation position when the target sequence and pooled probes are maximally aligned for complementarity).

The three cells are assigned probe pools that are perfectly complementary to the target except at the pooled position, which is occupied by a different pooled nucleotide in each probe.

With 3 pooled probes, all 4 possible single base pair states (wild and 3 mutants) are detected. A pool hybridizes with a target if some probe contained within that pool is complementary to that target.

A cell containing a pair (or more) of oligonucleotides lights up when a target complementary to any of the oligonucleotide in the cell is present. Using the simple strategy, each of the four possible targets (wild and three mutants) yields a unique hybridization pattern among the three cells.

Since a different pattern of hybridizing pools is obtained for each possible nucleotide in the target sequence corresponding to the pooled interrogation position in the probes, the identity of the nucleotide can be determined from the hybridization pattern of the pools. Whereas, a standard tiling requires four cells to detect and identify the possible single-base substitutions at one location, this simple pooled 45 strategy only requires three cells.

A more efficient pooling strategy for sequence analysis is the 'Trellis' strategy. In this strategy, each pooled probe has a segment of perfect complementarity to a reference sequence except at three pooled positions. One pooled position is an N pool. The three pooled positions may or may not be contiguous in a probe. The other two pooled positions are selected from the group of three pools consisting of (1) M or K, (2) R or Y and (3) W or S, where the single letters are IUPAC standard ambiguity codes. The sequence of a pooled probe is thus, of the form XXXN[(M/K) or (R/Y) or (W/S)][(M/K) or (R/Y) or (W/S)]XXXXX, where XXX represents bases complementary to the reference sequence. The three pooled positions may be in any order, and may be contiguous or separated by intervening nucleotides. For, the two positions occupied by [(M/K) or (R/Y) or (W/S)], two choices must be made. First, one must select one of the following three pairs of pooled nucleotides (1) M/K, (2) R/Y and (3) W/S. The one of three pooled nucleotides selected may be the same or different at the two pooled positions. Second, supposing, for example, one selects M/K at one position, one must then chose between M or K. This choice should result in selection of a pooled nucleotide comprising a nucleotide that complements the corresponding nucleotide in a reference sequence, when the probe and reference sequence are maximally aligned. The same principle governs the selection between R and Y, and between W and S. A trellis pool probe has one pooled position with four possibilities, and two pooled positions, each with two possibilities. Thus, a trellis pool probe comprises a mixture of 16 (4 x 2 x 2) probes. Since each pooled position includes one nucleotide that complements the corresponding nucleotide from the reference sequence, one of these 16 probes has a segment that is the exact complement of the reference sequence. A target sequence that is the same as the reference sequence (i.e., a wildtype target) gives a hybridization signal to each probe cell. Here, as in other tiling methods, the segment of complementarity should be sufficiently long to permit specific hybridization of a pooled probe to a reference sequence be detected relative to a variant of that reference sequence. Typically, the segment of complementarity is about 9-21 nucleotides.

A target sequence is analyzed by comparing hybridization intensities at three pooled probes, each having the structure described above. The segments complementary to the reference sequence present in the three pooled probes show some overlap.

Sometimes the segments are identical (other than at the interrogation positions). However, this need not be the case.

For example, the segments can tile across a reference sequence in increments of one nucleotide (i.e., one pooled probe differs from the next by the acquisition of one nucleotide at the 5□ end and loss of a nucleotide at the 3□ end). The three interrogation positions may or may not occur at the same relative positions within each pooled probe (i.e., spacing from a probe terminus). All that is required is that one of the three interrogation positions from each of the three pooled probes aligns with the same nucleotide in the reference sequence, and that this interrogation position is occupied by a different pooled nucleotide in each of the three probes. In one of the three probes, the interrogation position is occupied by an N. In the other two pooled probes the interrogation position is occupied by one of (M/K) or (R/Y) or (W/S).

In the simplest form of the trellis strategy, three pooled probes are used to analyze a single nucleotide in the reference sequence. Much greater economy of probes is achieved when more pooled probes are included in an array.

For example, consider an array of five pooled probes each having the general structure outlined above. Three of these pooled probes have an interrogation position that aligns with the same nucleotide in the reference sequence and are used to read that nucleotide. A different combination of three probes have an interrogation position that aligns with a different nucleotide in the reference sequence. Comparison of these three probe intensities allows analysis of this second nucleotide. Still another combination of three pooled probes from the set of five have an interrogation position that aligns with a third nucleotide in the reference sequence and these probes are used to analyze that nucleotide. Thus, three nucleotides in the reference sequence are fully analyzed from only five pooled probes. By comparison, the basic tiling strategy would require 12 probes for a similar analysis.

The trellis strategy employs an array of probes having at least three cells, each of which is occupied by a pooled probe as described above.

Consider the use of three such pooled probes for analyzing a target sequence, of which one position may contain any single base substitution to the reference sequence (i.e, there are four possible target sequences to be distinguished).

Three cells are occupied by pooled probes having a pooled interrogation position corresponding to the position of possible substitution in the target sequence, one cell with an □N□, one cell with one of □M□ or □K□, and one cell with □R□ or □Y□. An interrogation position corresponds to a nucleotide in the target sequence if it aligns adjacent with that nucleotide when the probe and target sequence are aligned to maximize complementarity. Note that although each of the pooled probes has two other pooled positions, these positions are not relevant for the present illustration. The positions are only relevant when more than one position in the target sequence is to be read, a circumstance that will be considered later. For present purposes, the cell with the □N□ in the interrogation position lights up for the wildtype sequence and any of the three single base substitutions of the target sequence.

A further class of strategies involving pooled probes are termed coding strategies. These strategies assign code words from some set of numbers to variants of a reference sequence.

Any number of variants can be coded. The variants can include multiple closely spaced substitutions, deletions or insertions. The designation letters or other symbols assigned to each variant may be any arbitrary set of numbers, in any order. For example, a binary code is often used, but codes to other bases are entirely feasible. The numbers are often assigned such that each variant has a designation having at least one digit and at least one nonzero value for that digit.

For example, in a binary system, a variant assigned the number 101, has a designation of three digits, with one possible nonzero value for each digit.

The designation of the variants are coded into an array of pooled probes comprising a pooled probe for each nonzero value of each digit in the numbers assigned to the variants.

For example, if the variants are assigned successive number in a numbering system of base m, and the highest number assigned to a variant has n digits, the array would have about n x (m -1) pooled probes. In general, $\log_m$ (3N+1) probes are required to analyze all variants of N locations in a reference sequence, each having three possible mutant substitutions.

For example, 10 base pairs of sequence may be analyzed with only 5 pooled probes using a binary coding system.

Each pooled probe has a segment exactly complementary to the reference sequence except that certain positions are pooled.

The segment should be sufficiently long to allow specific hybridization of the pooled probe to the reference sequence relative to a mutated form of the reference sequence. As in other tiling strategies, segments lengths of 9-21 nucleotides are typical. Often the probe has no nucleotides other than the 9-21 nucleotide segment. The pooled positions comprise nucleotides that allow the pooled probe to hybridize to every variant assigned a particular nonzero value in a particular digit. Usually, the pooled positions further comprises a nucleotide that allows the pooled probe to hybridize to the reference sequence. Thus, a wildtype target (or reference sequence) is immediately recognizable from all the pooled probes being lit.

When a target is hybridized to the pools, only those pools comprising a component probe having a segment that is exactly complementary to the target light up. The identity of the target is then decoded from the pattern of hybridizing pools. Each pool that lights up is correlated with a particular value in a particular digit. Thus, the aggregate hybridization patterns of each lighting pool reveal the value of each digit in the code defining the identity of the target hybridized to the array.

1.2.1.1.5.9. Bridging Strategy

Probes that contain partial matches to two separate (i.e., non contiguous) subsequences of a target sequence sometimes hybridize strongly to the target sequence. In certain instances, such probes have generated stronger signals than probes of the same length which are perfect matches to the target sequence. It is believed (but not necessary to the invention) that this observation results from interactions of a single target sequence with two or more probes simultaneously. This invention exploits this observation to provide arrays of probes having at least first and second segments, which are respectively complementary to first and second subsequences of a reference sequence. Optionally, the probes may have a third or more complementary segments. These probes can be employed in any of the strategies noted above.

The two segments of such a probe can be complementary to disjoint subsequences of the reference sequences or contiguous subsequences. * If the latter, the two segments in the probe are inverted relative to the order of the complement of the reference sequence. The two subsequences of the reference sequence each typically comprises about 3 to 30 contiguous nucleotides. The subsequences of the reference sequence are sometimes separated by 0, 1, 2 or 3 bases. Often the sequences, are adjacent and nonoverlapping.

The bridging strategy offers the following advantages:

(1) Higher discrimination between matched and mismatched probes, (2) The possibility of using longer probes in a bridging tiling, thereby increasing the specificity of the hybridization, without sacrificing discrimination, (3) The use of probes in which an interrogation position is located very off-center relative to the regions of target complementarity. This may be of particular advantage when, for example, when a probe centered about one region of the target gives low hybridization signal. The low signal is overcome by using a probe centered about an adjoining region giving a higher

hybridization signal. (4) Disruption of secondary structure that might result in annealing of certain probes (see previous discussion of helper mutations).

1.2.1.1.5.10. Deletion Tiling

Deletion tiling is related to both the bridging and helper mutant strategies described above. In the deletion strategy, comparisons are performed between probes sharing a common deletion but differing from each other at an interrogation position located outside the deletion. For example, a first probe comprises first and second segments, each exactly complementary to respective first and second subsequences of a reference sequence, wherein the first and second subsequences of the reference sequence are separated by a short distance (e.g., 1 or 2 nucleotides). The order of the first and second segments in the probe is usually the same as that of the complement to the first and second subsequences in the reference sequence.

Such tilings sometimes offer superior discrimination in hybridization intensities between the probe having an interrogation position complementary to the target and other probes. Thermodynamically, the difference between the hybridizations to matched and mismatched targets for the probe set shown above is the difference between a single-base bulge, and a large asymmetric loop (e.g., two bases of target, one of probe). This often results in a larger difference in stability than the comparison of a perfectly matched probe with a probe showing a single base mismatch in the basic tiling strategy.

The use of deletion or bridging probes is quite general. These probes can be used in any of the tiling strategies of the invention. As well as offering superior discrimination, the use of deletion or bridging strategies is advantageous for certain probes to avoid self-hybridization (either within a probe or between two probes of the same sequence)

1.2.1.1.6. Preparation of Target Samples

The target polynucleotide, whose sequence is to be determined, is usually isolated from a tissue sample. If the target is genomic, the sample may be from any tissue (except exclusively red blood cells). For example, whole blood, peripheral blood lymphocytes or PBMC, skin, hair or semen are convenient sources of clinical samples. These sources are also suitable if the target is RNA. Blood and other body fluids are also a convenient source for isolating viral nucleic acids. If the target is mRNA, the sample is obtained from a tissue in which the mRNA is expressed. If the polynucleotide in the sample is RNA, it is usually reverse transcribed to DNA. DNA samples or cDNA resulting from reverse transcription are usually amplified, e.g., by PCR. Depending on the selection of primers and amplifying enzyme(s), the amplification product can be RNA or DNA.

Paired primers are selected to flank the borders of a target polynucleotide of interest. More than one target can be simultaneously amplified by multiplex PCR in which multiple paired primers are employed. The target can be labelled at one or more nucleotides during or after amplification. For some target polynucleotides (depending on size of sample), e.g., episomal DNA, sufficient DNA is present in the tissue sample to dispense with the amplification step.

When the target strand is prepared in single-stranded form as in preparation of target RNA, the sense of the strand should of course be complementary to that of the probes on the chip. This is achieved by appropriate selection of primers.

The target is preferably fragmented before application to the chip to reduce or eliminate the formation of secondary structures in the target. The average size of targets segments following hybridization is usually larger than the size of probe on the chip.

1.2.1.2. Sequencing

This invention provides that the method of performing whole cell engineering may comprise the step of cell screening. In a preferred embodiment, this invention provides that the step of cell screening may comprise the step of genomic sequencing. In one exemplification, genome sequencing can be accomplished according to the enzymatic/Sanger method (described in F. Sanger, S. Nicklen, and A. R. Coulson, Proc. Nati. Acad. Sci, USA, 74:5463-5467 (1977)) and involve cloning and subcloning (described in U.S. Patent No. 4725677; Chen and Seeburg, DNA 4, 165-170 (1985); Lim et al., Gene Anal., Techn. 5, 32-39 (1988); PCR Protocols- A Guide to Methods and Applications. Innis et al., editors, Academic Press, San Diego (1990); Innis et al., Proc. Nat. Acad. Sci. USA 85, 9436-9440 (1988)).

In another exemplification, sequencing can be accomplished according to the chemical/Maxam and Gilbert method which is described in references: A. M. Maxam, and W. Gilbert, Proc. Nat. Acad. of Sci., USA, 74:560-564 (1977) and Church et al., Proc. Natl. Acad. Sci., 81:1991 (1984). In additional exemplifications, genome sequencing can be accomplished by methodology described by Guo and Wu (Guo and Wu, Nucleic Acids Res., 10:2065 (1982); and Meth. Enz.,100:60 (1983)) or those methods that utilize

3'hydroxy-protected and labeled nucleotides as exemplified in the following references: Churchich, J.E., Eur. J. Biochem., 231:736 (1995); Metzket, M.L.et al.,Nucleic Acids Research, 22:4259 (1994); Beabealashvilli, R.S. et al, Biochimica et Biophysica Acta, 868:136 (1986); Chidgeavadze, Z.G.; Kukhanova, M.K. et al.Biochimica et Biophysica Acta, 868:145 (1986); Hiratsuka, T et Biophysica Acta, 742:496 (1983); Jeng, S.J. and Guillory, R.J. J., Supramolecular Structure, 3:448 (1975).

The invention also provides that sequencing may be read by autoradiography using radioisotopes (as described in Ornstein et al., Biotechniques 2, 476 (1985)) or by using non-radioactively labeling strategies that have been integrated into partly automated DNA sequencing procedures (Smith et al., Nature M, 674-679 (1986) and EPO Patent No. 873 00998.9; Du Pont De Nemours EPO Application No. 03 59225; Ansorge et al., L Biochem. Biophys. Method 13, 325-32 (19860; Prober et al. Science M, 336-41 (1987); Applied Biosystems, PCT Application WO 91/05060; Smith et al., Science 235, G89 (1987); U.S. Patent Nos. 570973 and 689013), Du Pont De Nemours, U.S. Patents Nos. 881372 and 57566, Ansorge et al. Nucleic Acids Res. 15-, 4593-4602 (1987) and EMBL Patent Application DE P3724442 and P3805808.1) and Hitachi (JP 1-90844 and DE 4011991 Al; U.S. Patent No. 4,729,947; PCT Application W092/02635; U.S. Patent No. 594676; Beck, O'Keefe, Coull and Köster, Nucleic Acids Res. 7, 5115- 5123 (1989) .L7 and Beck and Köster, Anal. Chem. 62 2258-2270 (1990); Church et al., Science 240, 185-188 (1988); Köster et al., Nucleic Acids Res. Symposium Ser. No. 24, 318-321 (1991), University of Utah, PCT Application No. WO 90/15883; Smith et al., Nature (1986) 321:674- 679; Orion-Yhtyma Oy, U.S. Patent No. 277643; M. Uhlen et al. Nucleic Acids Res. 16, 3025-38 (1988); Cemu Bioteknik, PCT Application No. WO 89/09282 and Medical Research Council, GB, PCT Application No. WO 92/03575; Du Pont De Nemours, PCT Application WO 91/11533).

In addition, this invention provides for various methods of reading sequencing data such as capillary zone electrophoresis (described in Jorgenson et al., J. Chromatography 352, 337 (1986); Gesteland et al., Nucleic Acids Res. 18, 1415-1419 (1990)), mass spectrometry (including ES [described in Fenn et al. J. Phys. Chem. 18, 4451-59 (1984); PCT Application No. WO 90/14148; R.D. Smith et al., Anal. Chem. 62, 882-89 (1990) and B. Ardrey, Electrospray Mass Spectrometry, Spectroscopy Europe 4, 10-18 (1992)] and MALDI [Hillenkamp et al. Matrix Assisted UV-Laser Desorption/Ionization: A New

Approach to Mass Spectrometry of Large Biomolecules, Biological Mass Spectrometry (Burlingame and McCloskey, editors), Elsevier Science Publishers, Amsterdam, pp. 49-60, (1990); Williams et al., Science, 246, 1585-87 (1989); Williams et al., Rapid Communications in Mass Spectrometry, 4, 348-351 (1990)]), tube gel electrophoresis and a mass analyzer to sequence (described in EPO Patent Applications No. 0360676 Al and 0360677). In order to analyze the sequencing data, this invention provides for the use of probes in large arrays (as described in PCT patent Publication No. 92/10588; U.S. Patent No. 5,143,854; U.S. Application Serial No. 07/805,727; U.S. Patent No. 5,202,231; PCT patent Publication No. 89/10977).

This invention provides that the method of performing whole cell engineering may comprise the step of cell screening which in a particular embodiment may include the method of DNA amplification. In a particular embodiment, this invention provides that DNA amplification. DNA can be amplified by a variety of procedures including cloning (Sambrook et at., Molecular Cloning : A Laboratory Manual., Cold Spring Harbor Laboratory Press, 1989), polymerase chain reaction (PCR) (C.R. Newton and A. Graham, PCF, BIOS Publishers, 1994; Bevan et al., "Sequencing of PCR-Amplified DNA" PCR Meth. App. 4:222 (1992)), ligase chain reaction (LCR) (F. Barany Proc. Natl. Acad Sci USA 88, 189-93 (1991), strand displacement amplification (SDA) (G. Terrance Walker et al., Nucleic Acids Res. 22, 2670-77 (1994)) and variations such as RT-PCR (Arens, M. Clin Microbiol Rev, 12(4):612-26 (1999)), allele-specific amplification (ASA) (Nichols, W.C. et al. Genomics. Oct;5(3):535-40(1989); Giffard, P.M. et al. Anal Biochem, ;292(2):207-15 (2001)).

In additional embodiments of this invention, it provides for additional sequencing methods (as described in Labeit et al., MA 5, 173-177 (1986); Amersham, PCT-Application GB86/00349; Eckstein et al., Nucleic Acids Res. 1~, 9947 (1988); Max-Planck- Geselischaft, DE 3930312 Al; Saiki, R. et al., Science 239:487-491 (1998); Sarkat, G. and Bolander Mark E., Semi Exponential Cycle Sequencing Nucleic Acids Research, 1995, Vol. 23, No. 7, p. 1269-1270).

This invention also provides for the following sequencing strategies: shotgun sequencing, transposon-mediated directed sequencing (Strathmann, M. et al. Proc Natl Acad Sci USA (1991) 88:1247- 1250), and large scale variations thereof (as exemplified in K. B. Mullis et al., U.S. Pat. Nos. 4,683,202; 7/1987; 435/91; and 4,683,195, 7/1987; 435/6).

According to alternative embodiments of this invention, the step of genomic sequencing may include constructing ordered clone maps of DNA sequencing (as described in sections of U.S. Patent Publication No. 5604100 and PCT Patent Publication No. WO9627025). This invention provides that the method of genome sequencing be achieved by various steps that may utilize modifications of certain methods mentioned above (described in the following patents: PCT Publication Nos. WO9737041, WO9742348, WO9627025, WO9831834, WO9500530, and WO9831833; US Patent Publication Nos.US5604100, US5670321, US5453247, US5994058, and US5354656).

1.2.1.3. Annotating

In one aspect this invention discloses the use of a relational database system for storing and manipulating biomolecular sequence information and storing and displaying genetic information, the database including genomic libraries for a plurality of types of organisms, the libraries having multiple genomic sequences, at least some of which represent open reading frames located along a contiguous sequence on each the plurality of organisms' genomes, and a user interface capable of receiving a selection of two or more of the genomic libraries for comparison and displaying the results of the comparison. Associated with the database is a software system that allows a user to determine the relative position of a selected gene sequence within a genome. The system allows execution of a method of displaying the genetic locus of a biomolecular sequence. The method involves providing a database including multiple biomolecular sequences, at least some of which represent open reading frames located along a contiguous sequence on an organism's genome. The system also provides a user interface capable of receiving a selection of one or more probe open reading frames for use in determining homologous matches between such probe open reading frame(s) and the open reading frames in the genomic libraries, and displaying the results of the determination. An open reading frame for the sequence is selected and displayed together with adjacent open reading frames located upstream and downstream in the relative positions in which they occur on the contiguous sequence.

Also disclosed is a relational database system for storing biomolecular sequence information in a manner that allows sequences to be catalogued and searched according to one or more protein function hierarchies. The hierarchies allow searches for sequences based upon a protein's biological function or molecular function. Also disclosed is a mechanism for automatically grouping new sequences into protein function hierarchies. This mechanism uses descriptive information obtained from "external hits" which are matches of stored sequences against gene sequences stored in an external database such as GenBank. The descriptive information provided with the external database is evaluated according to a specific algorithm and used to automatically group the external hits (or the sequences associated with the hits) in the categories. Ultimately, the biomolecular sequences stored in databases of this invention are provided with both descriptive information from the external hit and category information from a relevant hierarchy or hierarchies.

Disclosed is a relational database system for storing biomolecular sequence information in a manner that allows sequences to be catalogued and searched according to association with one or more projects for obtaining full-length biomolecular sequences from shorter sequences. The relational database has sequence records containing information identifying one or more projects to which each of the sequence records belong. Each project groups together one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. The computer system has a user interface allowing a user to selectively view information regarding one or more projects. The relational database also provides interfaces and methods for accessing and manipulating and analyzing project-based information.

Polymer sequences are assembled into bins. A first number of bins are populated with polymer sequences. The polymer sequences in each bin are assembled into one or more consensus sequences representative of the polymer sequences of the bin. The consensus sequences of the bins are compared to determine relationships, if any, between the consensus sequences of the bins. The bins are modified based on the relationships between the consensus sequences of the bins. The polymer sequences are reassembled in the modified bins to generate one or more modified consensus sequences for each bin representative of the modified bins. In another aspect of the invention, sequence similarities and dissimilarities are analyzed in a set of polymer sequences. Pairwise alignment data is generated for pairs of the polymer sequences. The pairwise alignment data defines regions of similarity between the pairs of polymer sequences with boundaries. Additional boundaries in particular polymer sequences are determined by applying at least one boundary from at least one pairwise alignment for one pair of polymer sequences to at least one other pairwise alignment for another pair of polymer sequences including one of the particular polymer sequences. Additional regions of similarity are generated based on the boundaries.

1.2.1.3.1. ANNOTATING - GENERAL METHODOLOGY

In one aspect this present invention relates generally to relational databases for storing and retrieving biological information. More particularly the invention relates to systems and methods for providing sequences of biological molecules in a relational format allowing retrieval in a client-server environment and for providing full-length cDNA sequences in a relational format allowing retrieval in a client-server environment.

Informatics is the study and application of computer and statistical techniques to the management of information. In genome projects, bioinformatics includes the development of methods to search databases quickly, to analyze nucleic acid sequence information, and to predict protein sequence, structure and function from DNA sequence data.

Increasingly, molecular biology is shifting from the laboratory bench to the computer desktop. Today's researchers require advanced quantitative analyses, database comparisons, and computational algorithms to explore the relationships between sequence and phenotype. Thus, by all accounts, researchers can not and will not be able to avoid using computer resources to explore gene expression, gene sequencing and molecular structure.

One use of bioinformatics involves studying an organism's genome to determine the sequence and placement of its genes and their relationship to other sequences and genes within the genome or to genes in other organisms. Another use of bioinformatics involves studying genes differentially or commonly expressed in different tissues or cell lines (e.g. normal and cancerous tissue).

Such information is of significant interest in biomedical and pharmaceutical research, for instance to assist in the evaluation of drug efficacy and resistance.

The sequence tag method involves generation of a large number (e.g., thousands) of Expressed Sequence Tags ("ESTs") from cDNA libraries (each produced from a different tissue or sample). ESTs are partial transcript sequences that may cover different parts of the cDNA(s) of a gene, depending on cloning and sequencing strategy. Each EST includes about 50 to 300 nucleotides. If it is assumed that the number of tags is proportional to the abundance of transcripts in the tissue or cell type used to make the cDNA library, then any variation in the relative frequency of those tags, stored in computer databases, can be used to detect the differential abundance and potentially the expression of the corresponding genes.

To make genomic and EST information manipulation easy to perform and understand, sophisticated computer database systems have been developed. In one database system, developed by Incyte Pharmaceuticals, Inc. of Palo Alto, CA, genomic sequence data and the abundance levels of mRNA species represented in a given sample is electronically recorded and annotated with information available from public sequence databases such as GenBank. Examples of such databases include GenBank (NCBI) and TIGR. The resulting information is stored in a relational database that may be employed to determine relationships between sequences and genes within and among genomes and establish a cDNA profile for a given tissue and to evaluate changes in gene expression caused by disease progression, pharmacological treatment, aging, etc.

In one database system, developed by Incyte Pharmaceuticals, Inc. of Palo Alto, Calif., abundance levels of mRNA species represented in a given sample are electronically recorded and annotated with information available from public sequence databases such as GenBank. The resulting information is stored in a relational database that may be employed to establish a cDNA profile for a given tissue and to evaluate changes in gene expression caused by disease progression, pharmacological treatment, aging, etc.

Genetic information for a number of organisms has been catalogued in computer databases. Genetic databases for organisms such as *Eschericia coli*, *Haemophilus influenzae*, *Mycoplasma genitalium*, and *Mycoplasma pneumoniae*, among others, are publicly available. At present, however, complete sequence data is available for relatively few species, and the ability to manipulate sequence data within and between species and databases is limited.

While genetic data processing and relational database systems such as those developed by Incyte Pharmaceuticals, Inc. provide great power and flexibility in analyzing genetic information and gene expression information, this area of technology is still in its infancy and further improvements in genetic data processing and relational database systems and their content will help accelerate biological research for numerous applications.

In genome projects, bioinformatics includes the development of methods to search databases quickly, to analyze nucleic acid sequence information, and to predict protein sequence and structure from DNA sequence data. Increasingly, molecular biology is shifting from the laboratory bench to the computer desktop. Advanced quantitative analyses, database comparisons, and computational algorithms are needed to explore the relationships between sequence and phenotype.

1.2.1.3.2. ANNOTATING – EXEMPLARY ASPECTS

The annotation methods of this invention include those described in PCT patent publication Nos. 98/26407, 98/26408, and 99/49403 and United States Patent Nos. 6,023,659 and 5,953,727 and are herein incorporated by reference in their entirety to the same extent as if each individual patent or patent application were specifically and individually indicated to be incorporated by reference in its entirety.

Thus, in one aspect, this present invention provides relational database systems for storing and analyzing biomolecular sequence information together with biological annotations detailing the source and interpretation the sequence data. The present invention provides a powerful database tool for drug development and other research and development purposes.

The present invention provides relational database systems for storing and analyzing biomolecular sequence information together with biological detailing the source and interpretation the sequence data. Disclosed is a relational database systems for storing and displaying genetic information.

Associated with the database is a software system the allows a user to determine the relative position of a selected gene sequence within a genome. The system allows execution of a method of displaying the genetic locus of a biomolecular sequence. The method involves providing a database including multiple biomolecular sequences, at least some of which represent open reading frames located along a contiguous sequence on an organism's genome. An open reading frame for the sequence is selected and displayed together with adjacent open reading frames located upstream and downstream in the relative positions in which they occur on the contiguous sequence.

The invention provides a method of displaying the genetic locus of a biomolecular sequence. The method involve providing a database including multiple biomolecular sequences, at least some of which represent open reading frames located along a contiguous sequence on an organism's genome. The method further involves identifying a selected open reading frame, and displaying the selected open reading frame together with adjacent open reading frames located upstream and downstream from the selected open reading frame.

The adjacent open reading frames and the selected open reading frame are displayed in the relative positions in which they occur on the contiguous sequence, textually and/or graphically. The method of the invention may be practiced with sequences from microbial organisms, and the sequences may include nucleic acid or protein sequences.

The invention also provides a computer system including a database having multiple biomolecular sequences, at least some of which represent open reading frames located along a contiguous sequence on an organism□s genome.

The computer system also includes a user interface capable of identifying a selected open reading frame, and displaying the selected open reading frame together with adjacent open reading frames located upstream and downstream from the selected open reading frame. The adjacent the open reading frames and the selected open reading frame are displayed in the relative positions in which they occur on the contiguous sequence. The user interface may also capable of detecting a scrolling command, and based upon the direction and magnitude of the scrolling command, identifying a new selected open reading frame from the contiguous sequence.

The invention further provides a computer program product comprising a computer-usable medium having computer-readable program code embodied thereon relating to a database including multiple biomolecular sequences, at least some of which represent open reading frames located along a contiguous sequence on an organism's genome. The computer program product includes computer-readable program code for identifying a selected open reading frame, and displaying the selected open reading frame together with adjacent open reading frames located upstream and downstream from the selected open reading frame. The adjacent open reading frames and the selected open reading frame are displayed in the relative positions in which they occur on the contiguous sequence.

Comparative Genomics is a feature of the database system of the present invention which allows a user to compare the sequence data of sets of different organism types. Comparative searches may be formulated in a number of ways using the Comparative Genomics feature. For example, genes common to a set of organisms may be identified through a "commonality" query, and genes unique to one of a set of organisms may be identified through a "subtraction" query.

Electronic Southern is a feature of the present database system which is useful for identifying genomic libraries in which a given gene or ORF exists.

A Southern analysis is a conventional molecular biology technique in which a nucleic acid of known sequence is used to identify matching (complementary) sequences in a sample of nucleic acid to be analyzed. Like their laboratory counterparts, Electronic Southerns according to the present invention may be used to locate homologous matches between a "probe" DNA sequence and a large number of DNA sequences in one or more libraries.

The present invention provides a method of comparing genetic complements of different types of organisms. The method involves providing a database having sequence libraries with multiple biomolecular sequences for different types of organisms, where at least some of the sequences represent open reading frames located along one or more contiguous sequences on each of the organisms' genomes. The method further involves receiving a selection of two or more of the sequence libraries for comparison, determining open reading frames common or unique to the selected sequence libraries, and displaying the results of the determination.

The invention also provides a method of comparing genomic complements of different types of organisms. The method involves providing a database having genomic sequence libraries with multiple biomolecular sequences for different types of organisms, where at least some of the sequences represent open reading frames located along one or more contiguous sequences on each of the organisms' genomes. The method further involves receiving a selection of two or more of the sequence libraries for comparison, determining sequences common or unique to the selected sequence libraries, and displaying the results of the determination.

The invention further provides a computer system including a database containing genomic libraries for different types of organisms, which libraries have multiple genomic sequences, at least some of which representing open reading frames located along one or more contiguous sequences on each the organisms' genomes. The system also includes a user interface capable of receiving a selection of two or more genomic libraries for comparison and displaying the results of the comparison.

Another aspect of the present invention provides a method of identifying libraries in which a given gene exists. The method involves providing a database including genomic libraries for one or more types of organisms. The libraries have multiple genomic sequences, at least some of which represent open reading frames located along one or more contiguous sequences on each the organisms' genomes. The method further involves receiving a selection of one or more probe sequences, determining homologous matches between the selected probe sequences and the sequences in the genomic libraries, and displaying the results of the determination.

The invention also provides a computer system including a database including genomic libraries for one or more types of organisms, which libraries have multiple genomic sequences, at least some of which represent open reading frames located along one or more contiguous sequences on each the organisms' genomes. The system also includes a user interface capable of receiving a selection of one or more probe sequences for use in determining homologous matches between one or more probe sequences and the sequences in the genomic libraries, and displaying the results of the determination.

Also provided is a computer program product including a computer- usable medium having computer-readable program code embodied thereon relating to a database including genomic libraries for one or more types of organisms. The libraries have multiple genomic sequences, at least some of which represent open reading frames located along one or more contiguous sequences on each the organisms' genomes. The computer program product includes computer-readable program code for providing, within a computing system, an interface for receiving a selection of two or more genomic libraries for comparison, determining sequences common or unique to the selected genomic libraries, and displaying the results of the determination.

Additionally provided is a computer program product including a computer-usable medium having computer-readable program code embodied thereon relating to a database including genomic libraries for one or more types of organisms. The libraries have multiple genomic sequences, at least some of which represent open reading frames located along one or more contiguous sequences on each the organisms' genomes. The computer program product includes computer-readable program code for providing, within a computing system, an interface for receiving a selection of one or more probe open reading frames, determining homologous matches between the probe sequences and the sequences in the genomic libraries, and displaying the results of the determination.

The invention further provides a method of presenting the genetic complement of an organism. The method involves providing a database including sequence libraries for a plurality of types of organisms, where the libraries have multiple biomolecular sequences, at least some of which represent open reading frames located along one or more contiguous sequences on each of the organisms' genomes. The method further involves receiving a selection of one of the sequence libraries, determining open reading frames within the selected sequence library, and displaying the results as one or more unique identifiers for groups of related opening reading frames.

The present invention provides relational database systems for storing biomolecular sequence information in a manner that allows sequences to be catalogued and searched according to one or more protein function hierarchies. The hierarchies are provided to allow carefully tailored searches for sequences based upon a protein's biological function or molecular function. To make this capability available in large sequence databases, the invention provides a mechanism for automatically grouping new sequences into protein function hierarchies. This mechanism takes advantage of descriptive information obtained from "external hits" which are matches of stored sequences against gene sequences stored in an external database such as GenBank. The descriptive information provided with GenBank is evaluated according to a specific algorithm and used to automatically group the external hits (or the sequences associated with the hits) in the categories. Ultimately, the biomolecular sequences stored in databases of this invention are provided with both descriptive information from the external hit and category information from a relevant hierarchy or hierarchies.

The invention provides a computer system having a database containing records pertaining to a plurality of biomolecular sequences. At least some of the biomolecular sequences are grouped into a first hierarchy of protein function categories, the protein function categories specifying biological functions of proteins corresponding to the biomolecular sequences and the first hierarchy. The hierarchy includes a first set of protein function categories specifying biological functions at a cellular level, and a second set of protein function categories specifying biological functions at a level above the cellular level. The computer system of the invention also includes a user interface allowing a user to selectively view information regarding the plurality of biomolecular sequences as it relates to the first hierarchy. The computer system may also include additional protein function categories based, for example, on molecular or enzymatic function of proteins. The biomolecular sequences may include nucleic acid or amino acid sequences. Some of said biomolecular sequences may be provided as part of one or more projects for obtaining full-length gene sequences from shorter sequences, and the database records may contain information about such projects.

The invention also provides a method of using a computer system to present information pertaining to a plurality of biomolecular sequence records stored in a database. The method involves displaying a list of the records or a field for entering information identifying one or more of the records, identifying one or more of the records that a user has selected from the list or field, matching the one or more selected records with one or more protein function categories from a first hierarchy of protein function categories into which at least some of the biomolecular sequence records are grouped, and displaying the one or more categories matching the one or more selected records. The protein function categories specify biological functions of proteins corresponding to the biomolecular sequences and the first hierarchy includes a first set of protein function categories specifying biological functions at a cellular level, and a second set of protein function categories specifying biological functions at a tissue level. The method may also involve matching the records against other protein function hierarchies, such as hierarchies based on molecular and/or enzymatic function, and displaying the results. At least some of the biomolecular sequences may be provided as part of one or more projects for obtaining full-length gene sequences from shorter sequences, and the database records may contain information about those projects.

Additionally, the invention provides a method of using a computer system to present information pertaining to a plurality of biomolecular sequence records stored in a database. The method involves displaying a list of one or more protein biological function categories from a first hierarchy of protein biological function categories into which at least some of the biomolecular sequence records are grouped, identifying one or more of the protein biological function categories that a user has selected from the list, matching the one or more selected protein biological function categories with one or more biomolecular sequence records which are grouped in the selected protein biological function categories, and displaying the one or more sequence records matching the one or more selected protein biological function categories. The protein biological function categories specify biological functions of proteins corresponding to the biomolecular sequences and the first hierarchy includes a first set of protein biological function categories specifying biological functions at a cellular level, and a second set of protein biological function categories specifying biological functions at a tissue level. The method may also involve matching the records against other protein function hierarchies, such as hierarchies based on molecular and/or enzymatic function, and displaying the results. At least some of the biomolecular sequences may be provided as part of one or more projects for obtaining full-length gene sequences from shorter sequences, and the database records may contain information about those projects.

Another aspect of the invention provides a database system having a plurality of internal records. The database includes a plurality of sequence records specifying biomolecular sequences, at least some of which records reference hits to an external database, which hits specify genes having sequences that at least partially match those of the biomolecular sequences. The database also includes a plurality of external hit records specifying the hits to the external database, and at least some of the records reference protein function hierarchy categories which specify at least one of biological functions of proteins or molecular functions of proteins. At least some of the biomolecular sequences may be provided as part of one or more projects for obtaining full-length gene sequences from shorter sequences, and the database records may contain information about those projects.

Further aspects of the present invention provide a method of using a computer system and a computer readable medium having program instructions to automatically categorize biomolecular sequence records into protein function categories in an internal database. The method and program involve receiving descriptive information about a biomolecular sequence in the internal database from a record in an external database pertaining to a gene having a sequence that at least partially matches that of the biomolecular sequence. Next, a determination is made whether the descriptive information contains one or more terms matching one or more keywords associated with a first protein function category, the keywords being terms consistent with a classification in the first protein function category. When at least one keyword is found to match a term in the descriptive information, a determination is made whether the descriptive information contains a term matching one or more anti- keywords associated with the first protein function category, the anti- keywords being terms inconsistent with a classification in the first protein function category. Then, the biomolecular sequence is grouped in the first protein function category when the descriptive information contains a term matching a keyword but contains no term matching an anti- keyword.

with reference to the drawings,

The present invention provides relational database systems for storing biomolecular sequence information in a manner that allows sequences to be catalogued and searched according to one or more characteristics. The sequence information of the database is generated by one or more "projects" which are concerned with identifying the full- length coding sequence of a gene (i.e., mRNA). The projects involve the extension of an initial sequenced portion of a clone of a gene of interest (e.g., an EST) by a variety of methods which use conventional molecular biological techniques, recently developed adaptations of these techniques, and certain novel database applications. Data accumulated in these projects may be provided to the database of the present invention throughout the course of the projects and may be available to database users (subscribers) throughout the course of these projects for research, product (i.e., drug) development, and other purposes.

In a preferred embodiment, the database of the present invention and its associated projects may provide sequence and related data in amounts and forms not previously available. The present invention preferably makes partial and full-length sequence information for a given gene available to a user both during the course of the data acquisition and once the full-length sequence of the gene has been elucidated. The database also preferably provides a variety of tools for analysis and manipulation of the data, including Northern analysis and Expression summaries. The present invention should permit more complete and accurate annotation_ of sequence data, as well as the study of relationships between genes of different tissues, systems or organisms, and ultimately detailed expression studies of full-length gene sequences.

The invention provides a computer system including a database having sequence records containing information identifying one or more projects to which each of the sequence records belong. Each project groups together one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. The computer system also has a user interface allowing a user to selectively view information regarding one or more projects. The biomolecular sequences may include nucleic acid or amino acid sequences. The user interface may allow users to view at least three levels of project information including a project information results level listing at least some of the projects in said database, a sequence information results level listing at least some of the sequences associated with a given project, and a sequence retrieval results level sequentially listing monomers which comprise a given sequence.

A method of using a computer system and a computer program product to present information pertaining to a plurality of sequence records stored in a database are also provided by the present invention. The sequence records contain information identifying one or more projects to which each of the sequence records belong. Each of the projects groups one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. The method and program involve providing an interface for entering query information relating to one or more projects, locating data corresponding to the entered query information, and displaying the data corresponding to the entered query information.

Additionally, the invention provides a method of using a computer system to present information pertaining to a plurality of sequence records stored in a database. The sequence records contains information identifying one or more projects to which each of the sequence records belong. Each of the projects groups one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. The method involves displaying a list of one or more project identifiers, determining which project identifier or identifiers from the list is selected by a user, then displaying a second list of one or more biomolecular sequence identifiers associated with the selected project identifier or identifiers, determining which sequence identifier or identifiers from the second list has been selected by a user, and displaying a third list of one or more sequences corresponding to the selected sequence identifier or identifiers. Following the display of the third list, a determination may be made whether and which sequence from the third list has been selected by a user. If a sequence is selected, a sequence alignment search of the selected sequence against other databased sequences may be initiated, and the results of the alignment search displayed.

For Electronic Northern analysis, the invention further provides a computer system including a database having sequence records containing information identifying one or more projects to which each of the sequence records belong, each of said projects grouping one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. The system also has a user interface capable of allowing a user to select one or more project identifiers or project member identifiers specifying one or more sequences to be compared with one or more cDNA sequence libraries, and displaying matches resulting from that comparison.

A method of using a computer system to present comparative information pertaining to a plurality of sequence records stored in a database is also provided by the present invention. The sequence records contain information identifying one or more projects to which each of the sequence records belong, each of the projects grouping one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. The method involves providing an interface capable of allowing a user to select one or more project identifiers or project member identifiers specifying one or more sequences, comparing the one or more specified sequences with one or more cDNA sequence libraries, and displaying matches resulting from the comparison.

In addition, for Expression analysis, the invention provides a computer system including a database having sequence records containing information identifying one or more projects to which each of the sequence records belong, each of the projects grouping one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. The system also has a user interface allowing a user to view expression information pertaining to the projects by selecting one or more expression categories for a query, and displaying the result of the query.

A method of using a computer system to view expression information pertaining to one or more projects, each of the projects grouping one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence, is also provided in accordance with the present invention. The computer system includes a database storing a plurality of sequence records, the sequence records containing information identifying one or more projects to which each of the sequence records belong. The method involves providing an interface which allows a user to select one or more expression categories as a query, locating projects belonging to the selected one or more expression categories, and displaying a list of located projects.

Finally, the present invention provides a computer system including a database having sequence records containing information identifying one or more projects to which each of the sequence records belong, each of the projects grouping one or more biomolecular sequences generated during work to obtain a full-length gene sequence from a shorter sequence. This computer system has a user interface allowing a user to selectively view information regarding said one or more projects and which displays information to a user in a format common to one or more other sequence databases. These and other features and advantages of the invention will be described in more detail below with reference to the drawings.

Polymer sequences are assembled into bins. A first number of bins are populated with polymer sequences. The polymer sequences in each bin are assembled into one or more consensus sequences representative of the polymer sequences of the bin. The consensus sequences of the bins are compared to determine relationships, if any, between the consensus sequences. The bins are modified based on the relationships between the consensus sequences. The polymer sequences are reassembled in the modified bins to generate one or more modified consensus sequences for each bin representative of the modified bins.

In another aspect of the invention, sequence similarities and dissimilarities are analyzed in a set of polymer sequences. Pairwise alignment data is generated for pairs of the polymer sequences. The pairwise alignment data defines regions of similarity between the pairs of polymer sequences with boundaries. Additional boundaries in particular polymer sequences are determined by applying at least one boundary from at least one pairwise alignment for one pair of polymer sequences to at least one other pairwise alignment for another pair of polymer sequences including one of the particular polymer sequences. Additional regions of similarity are generated based on the boundaries

1.2.1.3.3. ANNOTATING - PREFERRED EMBODIMENTS

Generally, the present invention provides an improved relational database for storing and manipulating genomic sequence information. While the invention is described in terms of a database optimized for microbial data, it is by no means so limited. The invention may be employed to investigate data from various sources. For example, the invention covers databases optimized for other sources of sequence data, such as animal sequences (e.g., human, primate, rodent, amphibian, insect, etc.), plant sequences and microbial sequences. In the following description, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be apparent, however, that the present invention may be practiced without limitation to some of the specific details presented herein.

Generally, the present invention provides an improved relational database for storing sequence information. The invention may be employed to investigate data from various sources. For example, it may catalogue animal sequences (e.g., human, primate, rodent, amphibian, insect, etc.), plant sequences, and microbial sequences.

1.3. Transcriptome analysis or RNA profiling

The characterization of RNA expression and transcript populations (the transcriptome) can be referred to as RNA profiling and/or expression profiling, utilizing high throughput techniques such as RNA differential displays and DNA microarrays. One potential method to characterize gene expression, SAGE (Serial Analysis of Gene Expression) utilizes combinatorial chemistry technology and short sequence tags in the screening of compound libraries. For further information see references: Burge, C.B. 2001. Chipping away at the transcriptome. Nat Genet, 27(3): 232-4; Hughes, T.R. and Shoemaker, D.D. 2001. DNA microarrays for expression profiling. Curr Opin Chem Biol, 5(1): 21-5; Yamamoto, M. et al. 2001. Use of serial analysis of gene expression (SAGE) technology. J Immunol Methods 250(1-2):45-66.

1.3.1 Screening and selecting nucleotides for protein binding

An embodiment of this invention provides for screening methods that include the user of recombinant and *in vitro* chemical synthesis methods. In these hybrid methods, cell-free enzymatic machinery is employed to accomplish the *in vitro* synthesis of the library members (i.e., peptides or polynucleotides). In one type of method, RNA molecules with the ability to bind a predetermined protein or a predetermined dye molecule were selected by alternate rounds of selection and PCR amplification (**Tuerk and Gold**, 1990; **Ellington and Szostak**, 1990). A similar technique was used to identify DNA sequences which bind a predetermined human transcription factor (**Thiesen and Bach**, 1990; **Beaudry and Joyce**, 1992; PCT patent publications **WO 92/05258** and **WO 92/14843**).

1.4. Proteomics

In another embodiment of this invention, this invention relates to the emerging field of **proteomics**, Proteomics involves the qualitative and quantitative measurement of gene activity by detecting and quantitating expression at the protein level, rather than at the messenger RNA level. Proteomics also involves the study of non-genome encoded events, including the post-translational modification of proteins (including glycosylation or other modifications), interactions between proteins, and the location of proteins within a cell. The structure, function, and/or level of activity of the proteins expressed by the cell are also of interest. Essentially, proteomics involves the study of part or all of the status of the total protein contained within or secreted by a cell. Proteomics requires means of separating proteins in complex mixtures and identifying both low-and high-abundance species. Examples of powerful methods currently used to resolve complex protein mixtures are 2D gel electrophoresis, reverse phase HPLC, capillary electrophoresis, isoelectric focusing and related hybrid techniques. Commonly used protein identification techniques include N-terminal Edman and mass spectrometry (electrospray [ESI] or matrix-assisted laser desorption ionization [MALDI] MS) and sophisticated database search programs, such as SEQUEST, to identify proteins in World Wide Web protein and nucleic acid databases from the MS-MS spectra of their peptides. Using a computer, the output of the mass spectrometry can be analyzed so as to link a gene and the particular protein for which it codes. This overall process is sometimes referred to as "functional genomics".

For general information on proteome research, see, for example, J.S. Fruton, 1999, Proteins, Enzymes, Genes: The Interplay of Chemistry and Biology, Yale Univ. Pr.; Wilkins et al., 1997, Proteome Research: New Frontiers in Functional Genomics (Principles and Practice), Springer Verlag; A.J. Link, 1999, 2-D Proteome Analysis Protocols (Methods in Molecular Biology, 112, Humana Pr.); and Kamp et al., 1999, Proteome and Protein Analysis, Springer Verlag. Signal Transduction
See also, James, Peter, "Protein identification in the post-genome era: the rapid rise of proteomics", Q. Rev. Biophysics, Vol. 30, No. 4, pp. 279-331 (1997), which is incorporated by reference, herein.

1.4.1 Screening peptides: Peptide Display Methods

The present invention is further directed to a method for generating a selected mutant polynucleotide sequence (or a population of selected polynucleotide sequences) typically in the form of amplified and/or cloned polynucleotides, whereby the selected polynucleotide sequences(s) possess at least one desired phenotypic characteristic (e.g., encodes a polypeptide, promotes transcription of linked polynucleotides, binds a protein, and the like) which can be selected for. One method for identifying hybrid polypeptides that possess a desired structure or functional property, such as binding to a predetermined biological macromolecule (e.g., a receptor), involves the screening of a large library of polypeptides for individual library members which possess the desired structure or functional property conferred by the amino acid sequence of the polypeptide.

One method of screening peptides involves the display of a peptide sequence, antibody, or other protein on the surface of a bacteriophage particle or cell. Generally, in these methods each bacteriophage particle or cell serves as an individual library member displaying a single species of displayed peptide in addition to the natural bacteriophage or cell protein sequences. Each bacteriophage or cell contains the nucleotide sequence information encoding the particular displayed peptide sequence; thus, the displayed peptide sequence can be ascertained by nucleotide sequence determination of an isolated library member.

A well-known peptide display method involves the presentation of a peptide sequence on the surface of a filamentous bacteriophage, typically as a fusion with a bacteriophage coat protein. The bacteriophage library can be incubated with an immobilized, predetermined macromolecule or small molecule (e.g., a receptor) so that bacteriophage particles which present a peptide sequence that binds to the immobilized macromolecule can be differentially partitioned from those that do not present peptide sequences that bind to the predetermined macromolecule. The bacteriophage particles (i.e., library members) which are bound to the immobilized macromolecule are then recovered and replicated to amplify the selected bacteriophage sub-population for a subsequent round of affinity enrichment and phage replication. After several rounds of affinity enrichment and phage replication, the bacteriophage library members that are thus selected are isolated and the nucleotide sequence encoding the displayed peptide sequence is determined, thereby identifying the sequence(s) of peptides that bind to the predetermined macromolecule (e.g., receptor). Such methods are further described in PCT patent publications **WO 91/17271, WO 91/18980, WO 91/19818** and **WO 93/08278**.

The latter PCT publication describes a recombinant DNA method for the display of peptide ligands that involves the production of a library of fusion proteins with each fusion protein composed of a first polypeptide portion, typically comprising a variable sequence, that is available for potential binding to a predetermined macromolecule, and a second polypeptide portion that binds to DNA, such as the DNA vector encoding the individual fusion protein. When transformed host cells are cultured under conditions that allow for expression of the fusion protein, the fusion protein binds to the DNA vector encoding it. Upon lysis of the host cell, the fusion protein/vector DNA complexes can be screened against a predetermined macromolecule in much the same way as bacteriophage particles are screened in the phage-based display system, with the replication and sequencing of the DNA vectors in the selected fusion protein/vector DNA complexes serving as the basis for identification of the selected library peptide sequence(s).

The displayed peptide sequences can be of varying lengths, typically from 3-5000 amino acids long or longer, frequently from 5-100 amino acids long, and often from about 8-15 amino acids long. A library can comprise library members having varying lengths of displayed peptide sequence, or may comprise library members having a fixed length of displayed peptide sequence. Portions or all of the displayed peptide sequence(s) can be random, pseudorandom, defined set kernal, fixed, or the like. The present display methods include methods for *in vitro* and *in vivo* display of single-chain antibodies, such as nascent scFv on polysomes or scfv displayed on phage, which enable large-scale screening of scfv libraries having broad diversity of variable region sequences and binding specificities.

The present invention also provides random, pseudorandom, and defined sequence framework peptide libraries and methods for generating and screening those libraries to identify useful compounds (e.g., peptides, including single-chain antibodies) that bind to receptor molecules or epitopes of interest or gene products that modify peptides or RNA in a desired fashion. The random, pseudorandom, and defined sequence framework peptides are produced from libraries of peptide library members that comprise displayed peptides or displayed single-chain antibodies attached to a polynucleotide template from which the displayed peptide was synthesized. The mode of attachment may vary according to the specific embodiment of the invention selected, and can include encapsulation in a phage particle or incorporation in a cell.

1.4.2. Screening that utilizes in vitro translation systems

An embodiment of this invention provides for the use of *in vitro* translation during the step of screening. In vitro translation has been used to synthesize proteins of interest and has been proposed as a method for generating large libraries of peptides. These methods, generally comprising stabilized polysome complexes, are described further in

PCT patent publications WO 88/08453, WO 90/05785, WO 90/07003, WO 91/02076, WO 91/05058, and WO 92/02536. Applicants have described methods in which library members comprise a fusion protein having a first polypeptide portion with DNA binding activity and a second polypeptide portion having the library member unique peptide sequence; such methods are suitable for use in cell-free *in vitro* selection formats, among others.

1.4.3. Affinity enrichment

An aspect of this invention provides for the use of affinity enrichment which allows a very large library of peptides and single-chain antibodies to be screened and the polynucleotide sequence encoding the desired peptide(s) or single-chain antibodies to be selected. The polynucleotide can then be isolated and shuffled to recombine combinatorially the amino acid sequence of the selected peptide(s) (or predetermined portions thereof) or single-chain antibodies (or just VHI, VLI or CDR portions thereof). Using these methods, one can identify a peptide or single-chain antibody as having a desired binding affinity for a molecule and can exploit the process of shuffling to converge rapidly to a desired high-affinity peptide or scfv. The peptide or antibody can then be synthesized in bulk by conventional means for any suitable use (e.g., as a therapeutic or diagnostic agent).

A significant advantage of the present invention is that no prior information regarding an expected ligand structure is required to isolate peptide ligands or antibodies of interest. The peptide identified can have biological activity, which is meant to include at least specific binding affinity for a selected receptor molecule and, in some instances, will further include the ability to block the binding of other compounds, to stimulate or inhibit metabolic pathways, to act as a signal or messenger, to stimulate or inhibit cellular activity, and the like.

The present invention also provides a method for shuffling a pool of polynucleotide sequences selected by affinity screening a library of polysomes displaying nascent peptides (including single-chain antibodies) for library members which bind to a predetermined receptor (e.g., a mammalian proteinaceous receptor such as, for example, a peptidergic hormone receptor, a cell surface receptor, an intracellular protein which binds to other protein(s) to form intracellular protein complexes such as hetero-dimers and the like) or epitope (e.g., an immobilized protein, glycoprotein, oligosaccharide, and the like).

The invention also provides peptide libraries comprising a plurality of individual library members of the invention, wherein (1) each individual library member of said plurality comprises a sequence produced by shuffling of a pool of selected sequences, and (2) each individual library member comprises a variable peptide segment sequence or single-chain antibody segment sequence which is distinct from the variable peptide segment sequences or single-chain antibody sequences of other individual library members in said plurality (although some library members may be present in more than one copy per library due to uneven amplification, stochastic probability, or the like).

1.4.4. Antibody Display

The present method can be used to shuffle, by *in vitro* and/or *in vivo* recombination by any of the disclosed methods, and in any combination, polynucleotide sequences selected by antibody display methods, wherein an associated polynucleotide encodes a displayed antibody which is screened for a phenotype (e.g., for affinity for binding a predetermined antigen (ligand).

Various prokaryotic expression systems have been developed that can be manipulated to produce combinatorial antibody libraries which may be screened for high-affinity antibodies to specific antigens. Recent advances in the expression of antibodies in Escherichia coli and bacteriophage systems (see "alternative peptide display methods", *infra*) have raised the possibility that virtually any specificity can be obtained by either cloning antibody genes from characterized hybridomas or by de novo selection using antibody gene libraries (e.g., from Ig cDNA).

Combinatorial libraries of antibodies have been generated in bacteriophage lambda expression systems which may be screened as bacteriophage plaques or as colonies of lysogens (**Huse** et al, 1989); **Caton** and **Koprowski**, 1990; **Mullinax** et al, 1990; **Persson** et al, 1991). Various embodiments of bacteriophage antibody display libraries and lambda phage expression libraries have been described (**Kang** et al, 1991; **Clackson** et al, 1991; **McCafferty** et al, 1990; **Burton** et al, 1991; **Hoogenboom** et al, 1991; **Chang** et al, 1991; **Breitling** et al, 1991; **Marks** et al, 1991, p. 581; **Barbas** et al, 1992; **Hawkins** and **Winter**, 1992; **Marks** et al, 1992, p. 779; **Marks** et al, 1992, p. 16007; and **Lowman** et al, 1991; **Lerner** et al, 1992; all incorporated herein by reference). Typically, a bacteriophage antibody display library is screened with a receptor (e.g., polypeptide, carbohydrate, glycoprotein, nucleic acid) that is immobilized (e.g., by covalent linkage to a chromatography resin to enrich for reactive phage by affinity chromatography) and/or labeled (e.g., to screen plaque or colony lifts).

One particularly advantageous approach has been the use of so-called single-chain fragment variable (scfv) libraries (**Marks** et al, 1992, p. 779; **Winter** and **Milstein**, 1991; **Clackson** et al, 1991; **Marks** et al, 1991, p. 581; **Chaudhary** et al, 1990; **Chiswell** et al, 1992; **McCafferty** et al, 1990; and **Huston** et al, 1988). Various embodiments of scfv libraries displayed on bacteriophage coat proteins have been described. Bacteriophage display of scfv have already yielded a variety of useful antibodies and antibody fusion proteins. A bispecific single chain antibody has been shown to mediate efficient tumor cell lysis (**Gruber** et al, 1994). Intracellular expression of an anti-Rev scfv has been shown to inhibit HIV-1 virus replication *in vitro* (**Duan** et al, 1994), and intracellular expression of an anti-p21rar, scfv has been shown to inhibit meiotic maturation of Xenopus oocytes (**Biocca** et al, 1993). Recombinant scfv which can be used to diagnose HIV infection have also been reported, demonstrating the diagnostic utility of scfv (**Lilley** et al, 1994). Fusion proteins wherein an scFv is linked to a second polypeptide, such as a toxin or fibrinolytic activator protein, have also been reported (**Holvost** et al, 1992; **Nicholls** et al, 1993).

Various methods have been reported for increasing the combinatorial diversity of a scfv library to broaden the repertoire of binding species (idiotype spectrum). Enzymatic inverse PCR mutagenesis has been shown to be a simple and reliable method for constructing relatively large libraries of scfv site-directed hybrids (**Stemmer** et al, 1993), as has error-prone PCR and chemical mutagenesis (**Deng** et al, 1994). Riechmann (**Riechmann** et al, 1993) showed semi-rational design of an antibody scfv fragment using site-directed randomization by degenerate oligonucleotide PCR and subsequent phage display of the resultant scfv hybrids. Barbas (**Barbas** et al, 1992) attempted to circumvent the problem of limited repertoire sizes resulting from using biased variable region sequences by randomizing the sequence in a synthetic CDR region of a human tetanus toxoid-binding Fab.

Displayed peptide/polynucleotide complexes (library members) which encode a variable segment peptide sequence of interest or a single-chain antibody of interest are selected from the library by an affinity enrichment technique. This is accomplished by means of a immobilized macromolecule or epitope specific for the peptide sequence of interest, such as a receptor, other macromolecule, or other epitope species. Repeating the affinity selection procedure provides an enrichment of library members encoding the desired sequences, which may then be isolated for pooling and shuffling, for sequencing, and/or for further propagation and affinity enrichment.

The library members without the desired specificity are removed by washing. The degree and stringency of washing required will be determined for each peptide sequence or single-chain antibody of interest and the immobilized predetermined macromolecule or epitope. A certain degree of control can be exerted over the binding characteristics of the nascent peptide/DNA complexes recovered by adjusting the conditions of the binding incubation and the subsequent washing. The temperature, pH, ionic strength, divalent cations concentration, and the volume and duration of the washing will select for nascent peptide/DNA complexes within particular ranges of affinity for the immobilized macromolecule. Selection based on slow dissociation rate, which is usually predictive of high affinity, is often the most practical route. This may be done either by continued incubation in the presence of a saturating amount of free predetermined macromolecule, or by increasing the volume, number, and length of the washes. In each case, the rebinding of dissociated nascent peptide/DNA or peptide/RNA complex is prevented, and with increasing time, nascent peptide/DNA or peptide/RNA complexes of higher and higher affinity are recovered.

Additional modifications of the binding and washing procedures may be applied to find peptides with special characteristics. The affinities of some peptides are dependent on ionic strength or cation concentration. This is a useful characteristic for peptides that will be used in affinity purification of various proteins when gentle conditions for removing the protein from the peptides are required.

One variation involves the use of multiple binding targets (multiple epitope species, multiple receptor species), such that a scfv library can be simultaneously screened for a multiplicity of scfv which have different binding specificities. Given that the size of a scfv library often limits the diversity of potential scfv sequences, it is typically desirable to us scfv libraries of as large a size as possible. The time and economic considerations of generating a number of very large polysome scFv-display libraries can become prohibitive. To avoid this substantial problem, multiple predetermined epitope species (receptor species) can be concomitantly screened in a single library, or sequential screening against a number of epitope species can be used. In one variation, multiple target epitope species, each encoded on a separate bead (or subset of beads), can be mixed and incubated with a polysome-display scfv library under suitable binding conditions. The collection of beads, comprising multiple epitope species, can then be used to isolate, by affinity selection, scfv library members. Generally, subsequent affinity screening rounds can include the same mixture of beads, subsets thereof, or beads containing only one or two individual epitope species. This approach affords efficient screening, and is compatible with laboratory automation, batch processing, and high throughput screening methods.

1.4.5. Expression systems

The DNA expression constructs will typically include an expression control DNA sequence operably linked to the coding sequences, including naturally-associated or heterologous promoter regions. Preferably, the expression control sequences will be eukaryotic promoter systems in vectors capable of transforming or transfecting eukaryotic host cells. Once the vector has been incorporated into the appropriate host, the host is maintained under conditions suitable for high level expression of the nucleotide sequences, and the collection and purification of the mutant' "engineered" antibodies.

The DNA sequences will be expressed in hosts after the sequences have been operably linked to an expression control sequence (i.e., positioned to ensure the transcription and translation of the structural gene). These expression vectors are typically replicable in the host organisms either as episomes or as an integral part of the host chromosomal DNA. Commonly, expression vectors will contain selection markers, e.g., tetracycline or neomycin, to permit detection of those cells transformed with the desired DNA sequences (see, e.g., **USPN 4,704,362**, which is incorporated herein by reference).

In addition to eukaryotic microorganisms such as yeast, mammalian tissue cell culture may also be used to produce the polypeptides of the present invention (see **Winnacker**, 1987), which is incorporated herein by reference). Eukaryotic cells are actually preferred, because a number of suitable host cell lines capable of secreting intact immunoglobulins have been developed in the art, and include the CHO cell lines, various COS cell lines, HeLa cells, and myeloma cell lines, but preferably transformed Bcells or hybridomas. Expression vectors for these cells can include expression control sequences, such as an origin of replication, a promoter, an enhancer (**Queen** et al, 1986), and necessary processing information sites, such as ribosome binding sites, RNA splice sites, polyadenylation sites, and transcriptional terminator sequences. Preferred expression control sequences are promoters derived from immunoglobulin genes, cytomegalovirus, SV40, Adenovirus, Bovine Papilloma Virus, and the like.

Eukaryotic DNA transcription can be increased by inserting an enhancer sequence into the vector. Enhancers are cis-acting sequences of between 10 to 300 bp that increase transcription by a promoter. Enhancers can effectively increase transcription when either 5' or 3' to the transcription unit. They are also effective if located within an intron or within the coding sequence itself. Typically, viral enhancers are used, including SV40 enhancers, cytomegalovirus enhancers, polyoma enhancers, and adenovirus enhancers. Enhancer sequences from mammalian systems are also commonly used, such as the mouse immunoglobulin heavy chain enhancer.

Mammalian expression vector systems will also typically include a selectable marker gene. Examples of suitable markers include, the dihydrofolate reductase gene (DHFR), the thymidine kinase gene (TK), or prokaryotic genes conferring drug resistance. The first two marker genes prefer the use of mutant cell lines that lack the ability to grow without the addition of thymidine to the growth medium. Transformed cells can then be identified by their ability to grow on non-supplemented media. Examples of prokaryotic drug resistance genes useful as markers include genes conferring resistance to G418, mycophenolic acid and hygromycin.

The vectors containing the DNA segments of interest can be transferred into the host cell by well-known methods, depending on the type of cellular host. For example, calcium chloride transfection is commonly utilized for prokaryotic cells, whereas calcium phosphate treatment. lipofection, or electroporation may be used for other cellular hosts.

Other methods used to transform mammalian cells include the use of Polybrene, protoplast fusion, liposomes, electroporation, and micro-injection (see, generally, **Sambrook** et al, 1982 and 1989).

Once expressed, the antibodies, individual mutated immunoglobulin chains, mutated antibody fragments, and other immunoglobulin polypeptides of the invention can be purified according to standard procedures of the art, including ammonium sulfate precipitation, fraction column chromatography, gel electrophoresis and the like (see, generally, **Scopes**, 1982). Once purified, partially or to homogeneity as desired, the polypeptides may then be used therapeutically or in developing and performing assay procedures, immunofluorescent stainings, and the like (see, generally, **Lefkovits** and **Pernis**, 1979 and 1981; **Lefkovits**, 1997).

1.4.6 Two-Hybrid Based Screening Assays

This invention provides for screening a two-hybrid screening system to identify library members which bind a predetermined polypeptide sequence. The selected library members are pooled and shuffled by *in vitro* and/or *in vivo* recombination. The shuffled pool can then be screened in a yeast two hybrid system to select library members which bind said predetermined polypeptide sequence (e. g., and SH2 domain) or which bind an alternate predetermined polypeptide sequence (e.g., an SH2 domain from another protein species).

An approach to identifying polypeptide sequences which bind to a predetermined polypeptide sequence has been to use a so-called "two-hybrid" system wherein the predetermined polypeptide sequence is present in a fusion protein (**Chien** et al, 1991). This approach identifies protein-protein interactions *in vivo* through reconstitution of a transcriptional activator (**Fields and Song**, 1989), the yeast Gal4 transcription protein. Typically, the method is based on the properties of the yeast Gal4 protein, which consists of separable domains responsible for DNA-binding and transcriptional activation. Polynucleotides encoding two hybrid proteins, one consisting of the yeast Gal4 DNA-binding domain fused to a polypeptide sequence of a known protein and the other consisting of the Gal4 activation domain fused to a polypeptide sequence of a second protein, are constructed and introduced into a yeast host cell. Intermolecular binding between the two fusion proteins reconstitutes the Gal4 DNA-binding domain with the Gal4 activation domain, which leads to the transcriptional activation of a reporter gene (e.g., *lacz, HIS3)* which is operably linked to a Gal4 binding site. Typically, the two-hybrid method is used to identify novel polypeptide sequences which interact with a known protein (**Silver and Hunt**, 1993; **Durfee** et al, 1993; **Yang** et al, 1992; **Luban** et al, 1993; **Hardy** et al, 1992; **Bartel** et al, 1993; and **Vojtek** et al, 1993). However, variations of the two-hybrid method have been used to identify mutations of a known protein that affect its binding to a second known protein (**Li and Fields,** 1993; **Lalo** et al, 1993; **Jackson** et al, 1993; and **Madura** et al, 1993). Two-hybrid systems have also been used to identify interacting structural domains of two known proteins (**Bardwell** et al, 1993; **Chakrabarty** et al, 1992; **Staudinger** et al, 1993; and **Milne and Weaver** 1993) or domains responsible for oligomerization of a single protein (**Iwabuchi** et al, 1993; **Bogerd** et al, 1993). Variations of two-hybrid systems have been used to study the *in vivo* activity of a proteolytic enzyme (**Dasmahapatra** et al, 1992). Alternatively, an E. coli/BCCP interactive screening system (**Germino** et al, 1993; **Guarente**, 1993) can be used to identify interacting protein sequences (i.e., protein sequences which heterodimerize or form higher order heteromultimers). Sequences selected by a two-hybrid system can be pooled and shuffled and introduced into a two-hybrid system for one or more subsequent rounds of screening to identify polypeptide sequences which bind to the hybrid containing the predetermined binding sequence. The sequences thus identified can be compared to identify consensus sequence(s) and consensus sequence kernals.

1.4.7. Improved methods for cellular engineering, protein expression profiling, differential labeling of peptides, and novel reagents therefore

In one embodiment, this invention relates to peptide chemistry, proteomics, and mass spectrometry technology. In particular, the invention provides novel methods for determining polypeptide profiles and protein expression variations, as with proteome analyses. The present invention provides methods of simultaneously identifying and quantifying individual proteins in complex protein mixtures by selective differential labeling of amino acid residues followed by chromatographic and mass spectrographic analysis.

The diagnosis and treatment, as well as the predisposition of, a variety of diseases and disorders may often be accomplished through identification and quantitative measurement of polypeptide expression variations between different cell types and cell states. Biochemical pathways and metabolic networks can also be analyzed by globally and quantitatively measuring protein expression in various cell types and biological states (see, e.g., Ideker (2001) Science 292:929-934).

State-of-the-art techniques such as liquid-chromatography-electrospray-ionization tandem mass spectrometry have, in conjunction with database-searching computer algorithms, revolutionized the analysis of biochemical species from complex biological mixtures. With these techniques, it is now possible to perform high-throughput protein identification at picomolar to subpicomolar levels from complex mixtures of biological molecules (see, e.g., Dongre (1997) Trends Biotechnol. 15:418-425).
One such method is based on a class of chemical reagents termed isotope-coded affinity tags (ICATs) and tandem mass spectrometry. The method labels multiple cysteinyl residues and uses stable isotope dilution techniques. For example, Gygi (1999) Nat. Biotechnol. 10:994-999, compared protein expression in a yeast using ethanol or galactose as a carbon source. The measured differences in protein expression correlated with known yeast metabolic function under glucose-repressed conditions.

In another technique, two different protein mixtures for quantitative comparison are digested to peptide mixtures, the peptides mixtures are separately methylated using either d0- or d3-methanol, the mixtures of methylated peptide combined and subjected to microcapillary HPLC-MS/MS (see, e.g., Goodlett, D. R., et al., (2000) "Differential stable isotope labeling of peptides for quantitation and *de novo* sequence derivation," 49th ASMS; Zhou, H; Watts, JD; Aebersold, R. A systematic approach to the analysis of protein phosphorylation.; Comment In: Nat Biotechnol. 2001 Apr;19(4):317-8; Nature Biotechnology 2001 Apr, 19(4):375-8). Parent proteins of methylated peptides are identified by correlative database searching of fragment ion spectra using a computer program assisted paradigms or automated *de novo* sequencing that compares all tandem mass spectra of d0- and d3-methylated peptide ion pairs. In Goodlett (2000) supra, ratios of proteins in two different mixtures were calculated for d0- to d3-methylated peptide pairs. However, there are several limitations to this approach, including: use of differential labeling reagents, which relied on stable isotopes, which are expensive, and not flexible to differential labeling of more than two mixtures of peptides; labeling methods limited only to methylation of carboxy-termini; protein expression profiling limited to duplex comparison; one dimensional capillary HPLC chromatography was employed to separate peptides, which doesn't has enough capacity and resolving power for complex mixtures of peptides.

In one embodiment this invention provides a method for identifying proteins by differential labeling of peptides, the method comprising the following steps: (a) providing a sample comprising a polypeptide; (b) providing a plurality of labeling reagents which differ in molecular mass that can generate differential labeled peptides that do not differ in chromatographic retention properties and do not differ in ionization and detection properties in mass spectrographic analysis, wherein the differences in molecular mass are distinguishable by mass spectrographic analysis; (c) fragmenting the polypeptide into peptide fragments by enzymatic digestion or by non-enzymatic fragmentation; (d) contacting the labeling reagents of step (b) with the peptide fragments of step (c), thereby labeling the peptides with the differential labeling reagents; (e) separating the peptides by chromatography to generate an eluate; (f) feeding the eluate of step (e) into a mass spectrometer and quantifying the amount of each peptide and generating the sequence of each peptide by use of the mass spectrometer; (g) inputting the sequence to a computer program product which compares the inputted sequence to a database of polypeptide sequences to identify the polypeptide from which the sequenced peptide originated.

In one aspect, the sample of step (a) comprises a cell or a cell extract. The method can further comprise providing two or more samples comprising a polypeptide. One or more of the samples can be derived from a wild type cell and one sample can be derived from an abnormal or a modified cell. The abnormal cell can be a cancer cell. The modified cell can be a cell that is mutagenized &/or treated with a chemical, a physiological factor, or the presence of another organism (including,.e.g. a eukaryotic organism, prokaryotic organism, virus, vector, prion, or part thereof), &/or exposed to an environmental factor or change or physical force (including, e.g., sound, light, heat, sonication, and radiation). The modification can be genetic change (including, for example, a change in DNA or RNA sequence or content) or otherwise.

In one aspect, the method further comprises purifying or fractionating the polypeptide before the fragmenting of step (c). The method can further comprise purifying or fractionating the polypeptide before the labeling of step (d). The method can further comprise purifying or fractionating the labeled peptide before the chromatography of step (e). In alternative aspects, the purifying or fractionating comprises a method selected from the group consisting of size exclusion chromatography, size exclusion chromatography, HPLC, reverse phase HPLC and affinity purification. In one aspect, the method further comprises contacting the polypeptide with a labeling reagent of step (b) before the fragmenting of step (c).

In one aspect, the labeling reagent of step (b) comprises the general formulae selected from the group consisting of: Z^AOH and Z^BOH, to esterify peptide C-terminals and/or Glu and Asp side chains; Z^ANH_2 and Z^BNH_2, to form amide bond with peptide C-terminals and/or Glu and Asp side chains; and Z^ACO_2H and Z^BCO_2H. to form amide bond with peptide N-terminals and/or Lys and Arg side chains; wherein Z^A and Z^B independently of one another comprise the general formula $R\text{-}Z^1\text{-}A^1\text{-}Z^2\text{-}A^2\text{-}Z^3\text{-}A^3\text{-}Z^4\text{-}A^4\text{-}$, Z^1, Z^2, Z^3, and Z^4 independently of one another, are selected from the group consisting of nothing, O, OC(O), OC(S), OC(O)O, OC(O)NR, OC(S)NR, $OSiRR^1$, S, SC(O), SC(S), SS, S(O), $S(O_2)$, NR, NRR^{1+}, C(O), C(O)O, C(S), C(S)O, C(O)S, C(O)NR, C(S)NR, $SiRR^1$, $(Si(RR^1)O)_n$, $SnRR^1$, $Sn(RR^1)O$, $BR(OR^1)$, BRR^1, $B(OR)(OR^1)$, $OBR(OR^1)$, $OBRR^1$, and $OB(OR)(OR^1)$, and R and R^1 is an alkyl group, A^1, A^2, A^3, and A^4 independently of one another, are selected from the group consisting of nothing or $(CRR^1)_n$, wherein R, R^1, independently from other R and R^1 in Z^1 to Z^4 and independently from other R and R^1 in A^1 to A^4, are selected from the group consisting of a hydrogen atom, a halogen atom and an alkyl group; "n" in Z^1 to Z^4, independent of n in A^1 to A^4, is an integer having a value selected from the group consisting of 0 to about 51; 0 to about 41; 0 to about 31; 0 to about 21, 0 to about 11 and 0 to about 6.

In one aspect, the alkyl group (see definition below) is selected from the group consisting of an alkenyl, an alkynyl and an aryl group. One or more C-C bonds from $(CRR^1)_n$ can be replaced with a double or a triple bond; thus, in alternative aspects, an R or an R^1 group is deleted. The $(CRR^1)_n$ can be selected from the group consisting of an *o*-arylene, an *m*-arylene and a *p*-arylene, wherein each group has none or up to 6 substituents. The $(CRR^1)_n$ can be selected from the group consisting of a carbocyclic, a bicyclic and a tricyclic fragment, wherein the fragment has up to 8 atoms in the cycle with or without a heteroatom selected from the group consisting of an O atom, a N atom and an S atom.

In one aspect, two or more labeling reagents have the same structure but a different isotope composition. For example, in one aspect, Z^A has the same structure as Z^B, while Z^A has a different isotope composition than Z^B. In alternative aspects, the isotope is boron-10 and boron-11; carbon-12 and carbon-13; nitrogen-14 and nitrogen-15; and, sulfur-32 and sulfur-34. In one aspect, where the isotope with the lower mass is x and the isotope with the higher mass is y, and x and y are integers, x is greater than y.

In alternative aspects, x and y are between 1 and about 11, between 1 and about 21, between 1 and about 31, between 1 and about 41, or between 1 and about 51.

In one aspect, the labeling reagent of step (b) comprises the general formulae selected from the group consisting of: $CD_3(CD_2)_nOH$ / $CH_3(CH_2)_nOH$, to esterify peptide C-terminals, where n = 0, 1, 2 or y; $CD_3(CD_2)_nNH_2$ / $CH_3(CH_2)_nNH_2$, to form amide bond with peptide C-terminals, where n = 0, 1, 2 or y; and, $D(CD_2)_nCO_2H$ / $H(CH_2)_nCO_2H$, to form amide bond with peptide N-terminals, where n = 0, 1, 2 or y; wherein D is a deuteron atom, and y is an integer selected from the group consisting of about 51; about 41; about 31; about 21, about 11; about 6 and between about 5 and 51.

In one aspect, the labeling reagent of step (b) can comprise the general formulae selected from the group consisting of: Z^AOH and Z^BOH to esterify peptide C-terminals; Z^ANH_2 / Z^BNH_2 to form an amide bond with peptide C-terminals; and, Z^ACO_2H / Z^BCO_2H to form an amide bond with peptide N-terminals; wherein Z^A and Z^B have the general formula $R\text{-}Z^1\text{-}A^1\text{-}Z^2\text{-}A^2\text{-}Z^3\text{-}A^3\text{-}Z^4\text{-}A^4\text{-}$; Z^1, Z^2, Z^3, and Z^4, independently of one another, are selected from the group consisting of nothing, O, OC(O), OC(S), OC(O)O, OC(O)NR, OC(S)NR, $OSiRR^1$, S, SC(O), SC(S), SS, S(O), $S(O_2)$, NR, NRR^{1+}, C(O), C(O)O, C(S), C(S)O, C(O)S, C(O)NR, C(S)NR, $SiRR^1$, $(Si(RR^1)O)n$, $SnRR^1$, $Sn(RR^1)O$, $BR(OR^1)$, BRR^1, $B(OR)(OR^1)$, $OBR(OR^1)$, $OBRR^1$, and $OB(OR)(OR^1)$; A^1, A^2, A^3, and A^4, independently of one another, are selected from the group consisting of nothing and the general formulae $(CRR^1)n$, and, R and R^1 is an alkyl group.

In one aspect, a single C-C bond in a $(CRR^1)n$ group is replaced with a double or a triple bond; thus, the R and R^1 can be absent. The $(CRR^1)n$ can comprise a moiety selected from the group consisting of an *o*-arylene, an *m*-arylene and a *p*-arylene, wherein the group has none or up to 6 substituents. The group can comprise a carbocyclic, a bicyclic, or a tricyclic fragments with up to 8 atoms in the cycle, with or without a heteroatom selected from the group consisting of an O atom, an N atom and an S atom. In one aspect, R, R^1, independently from other R and R^1 in Z^1 - Z^4 and independently from other R and R^1 in A^1 - A^4, are selected from the group consisting of a hydrogen atom, a halogen and an alkyl group. The alkyl group (see definition below) can be an alkenyl, an alkynyl or an aryl group.

In one aspect, the "n" in Z^1 - Z^4 is independent of n in A^1 - A^4 and is an integer selected from the group consisting of about 51; about 41; about 31; about 21, about 11 and about 6. In one aspect, Z^A has the same structure a Z^B but Z^A further comprises *x* number of -CH_2- fragment(s) in one or more A^1 - A^4 fragments, wherein *x* is an integer. In one aspect, Z^A has the same structure a Z^B but Z^A further comprises *x* number of -CF_2- fragment(s) in one or more A^1 - A^4 fragments, wherein *x* is an integer. In one aspect, Z^A comprises *x* number of protons and Z^B comprises *y* number of halogens in the place of protons, wherein *x* and *y* are integers. In one aspect, Z^A contains *x* number of protons and Z^B contains *y* number of halogens, and there are *x* - *y* number of protons remaining in one or more A^1 - A^4 fragments, wherein *x* and *y* are integers. In one aspect, Z^A further comprises *x* number of –O- fragment(s) in one or more A^1 - A^4 fragments, wherein *x* is an integer. In one aspect, Z^A further comprises *x* number of –S- fragment(s) in one or more A^1 - A^4 fragments, wherein *x* is an integer. In one aspect, Z^A further comprises *x* number of –O- fragment(s) and Z^B further comprises *y* number of –S- fragment(s) in the place of –O- fragment(s), wherein *x* and *y* are integers. In one aspect, Z^A further comprises *x* - *y* number of –O- fragment(s) in one or more A^1 - A^4 fragments, wherein *x* and *y* are integers.

In alternative aspects, *x* and *y* are integers selected from the group consisting of between 1 about 51; between 1 about 41; between 1 about 31; between 1 about 21, between 1 about 11 and between 1 about 6, wherein *x* is greater than *y*.

In one aspect, the labeling reagent of step (b) comprises the general formulae selected from the group consisting of: $CH_3(CH_2)_nOH/CH_3(CH_2)_{n+m}OH$, to esterify peptide C-terminals, where n = 0, 1, 2, ..., y; m = 1, 2, ..., y; $CH_3(CH_2)_n NH_2$ / $CH_3(CH_2)_{n+m}NH_2$, to form amide bond with peptide C-terminals, where n = 0, 1, 2, ..., y; m = 1, 2, ..., y; and, $H(CH_2)_nCO_2H$ / $H(CH_2)_{n+m}CO_2H$, to form amide bond with peptide N-terminals, where n = 0, 1, 2, ..., y; m = 1, 2, ..., y; wherein n, m and y are integers. In one aspect, n, m and y are integers selected from the group consisting of about 51; about 41; about 31; about 21, about 11; about 6 and between about 5 and 51.

In one aspect, the separating of step (e) comprises a liquid chromatography system, such as a multidimensional liquid chromatography or a capillary chromatography system. In one aspect, the mass spectrometer comprises a tandem mass spectrometry device. In one aspect, the method further comprises quantifying the amount of each polypeptide or each peptide.

The invention provides a method for defining the expressed proteins associated with a given cellular state, the method comprising the following steps: (a) providing a sample comprising a cell in the desired cellular state; (b) providing a plurality of labeling reagents which differ in molecular mass that can generate differential labeled peptides that do not differ in chromatographic retention properties and do not differ in ionization and detection properties in mass spectrographic analysis, wherein the differences in molecular mass are distinguishable by mass spectrographic analysis; (c) fragmenting polypeptides derived from the cell into peptide fragments by enzymatic digestion or by non-enzymatic fragmentation; (d) contacting the labeling reagents of step (b) with the peptide fragments of step (c), thereby labeling the peptides with the differential labeling reagents; (e) separating the peptides by chromatography to generate an eluate; (f) feeding the eluate of step (e) into a mass spectrometer and quantifying the amount of each peptide and generating the sequence of each peptide by use of the mass spectrometer; (g) inputting the sequence to a computer program product which compares the inputted sequence to a database of polypeptide sequences to identify the polypeptide from which the sequenced peptide originated, thereby defining the expressed proteins associated with the cellular state.

The invention provides a method for quantifying changes in protein expression between at least two cellular states, the method comprising the following steps: (a) providing at least two samples comprising cells in a desired cellular state; (b) providing a plurality of labeling reagents which differ in molecular mass that can generate differential labeled peptides that do not differ in chromatographic retention properties and do not differ in ionization and detection properties in mass spectrographic analysis, wherein the differences in molecular mass are distinguishable by mass spectrographic analysis; (c) fragmenting polypeptides derived from the cells into peptide fragments by enzymatic digestion or by non-enzymatic fragmentation; (d) contacting the labeling reagents of step (b) with the peptide fragments of step (c), thereby labeling the peptides with the differential labeling reagents, wherein the labels used in one same are different from the labels used in other samples; (e) separating the peptides by chromatography to generate an eluate; (f) feeding the eluate of step (e) into a mass spectrometer and quantifying the amount of each peptide and generating the sequence of each peptide by use of the mass spectrometer; (g) inputting the sequence to a computer program product which identifies from which sample each peptide was derived, compares the inputted sequence to a database of polypeptide sequences to identify the polypeptide from which the sequenced peptide originated, and compares the amount of each polypeptide in each sample, thereby quantifying changes in protein expression between at least two cellular states.

The invention provides a method for identifying proteins by differential labeling of peptides, the method comprising the following steps: (a) providing a sample comprising a polypeptide; (b) providing a plurality of labeling reagents which differ in molecular mass but do not differ in chromatographic retention properties and do not differ in ionization and detection properties in mass spectrographic analysis, wherein the differences in molecular mass are distinguishable by mass spectrographic analysis; (c) fragmenting the polypeptide into peptide fragments by enzymatic digestion or by non-enzymatic fragmentation; (d) contacting the labeling reagents of step (b) with the peptide fragments of step (c), thereby labeling the peptides with the differential labeling reagents; (e) separating the peptides by multidimensional liquid chromatography to generate an eluate; (f) feeding the eluate of step (e) into a tandem mass spectrometer and quantifying the amount of each peptide and generating the sequence of each peptide by use of the mass spectrometer; (g) inputting the sequence to a computer program product which compares the inputted sequence to a database of polypeptide sequences to identify the polypeptide from which the sequenced peptide originated.

The invention provides a chimeric labeling reagent comprising (a) a first domain comprising a biotin; and (b) a second domain comprising a reactive group capable of covalently binding to an amino acid, wherein the chimeric labeling reagent comprises at least one isotope. The isotope(s) can be in the first domain or the second domain. For example, the isotope(s) can be in the biotin.

In alternative aspects, the isotope can be a deuterium isotope, a boron-10 or boron-11 isotope, a carbon-12 or a carbon-13 isotope, a nitrogen-14 or a nitrogen-15 isotope, or, a sulfur-32 or a sulfur-34 isotope. The chimeric labeling reagent can comprise two or more isotopes. The chimeric labeling reagent reactive group capable of covalently binding to an amino acid can be a succimide group, an isothiocyanate group or an isocyanate group. The reactive group can be capable of covalently binding to an amino acid binds to a lysine or a cysteine.

The chimeric labeling reagent can further comprising a linker moiety linking the biotin group and the reactive group. The linker moiety can comprise at least one isotope. In one aspect, the linker is a cleavable moiety that can be cleaved by, e.g., enzymatic digest or by reduction.

The invention provides a method of comparing relative protein concentrations in a sample comprising (a) providing a plurality of differential small molecule tags, wherein the small molecule tags are structurally identical but differ in their isotope composition, and the small molecules comprise reactive groups that covalently bind to cysteine or lysine residues or both; (b) providing at least two samples comprising polypeptides; (c) attaching covalently the differential small molecule tags to amino acids of the polypeptides; (d) determining the protein concentrations of each sample in a tandem mass spectrometer; and, (d) comparing relative protein concentrations of each sample. In one aspect, the sample comprises a complete or a fractionated cellular sample.

In one aspect of the method, the differential small molecule tags comprise a chimeric labeling reagent comprising (a) a first domain comprising a biotin; and, (b) a second domain comprising a reactive group capable of covalently binding to an amino acid, wherein the chimeric labeling reagent comprises at least one isotope. The isotope can be a deuterium isotope, a boron-10 or boron-11 isotope, a carbon-12 or a carbon-13 isotope, a nitrogen-14 or a nitrogen-15 isotope, or, a sulfur-32 or a sulfur-34 isotope. The chimeric labeling reagent can comprise two or more isotopes. The reactive group can be capable of covalently binding to an amino acid is selected from the group consisting of a succimide group, an isothiocyanate group and an isocyanate group.

The invention provides a method of comparing relative protein concentrations in a sample comprising (a) providing a plurality of differential small molecule tags, wherein the differential small molecule tags comprise a chimeric labeling reagent comprising (i) a first domain comprising a biotin; and, (ii) a second domain comprising a reactive group capable of covalently binding to an amino acid, wherein the chimeric labeling reagent comprises at least one isotope; (b) providing at least two samples comprising polypeptides; (c) attaching covalently the differential small molecule tags to amino acids of the polypeptides; (d) isolating the tagged polypeptides on a biotin-binding column by binding tagged polypeptides to the column, washing non-bound materials off the column, and eluting tagged polypeptides off the column; (e) determining the protein concentrations of each sample in a tandem mass spectrometer; and, (f) comparing relative protein concentrations of each sample. The details of one or more embodiments of the invention are set forth in the accompanying drawings and the description below. Other features, objects, and advantages of the invention will be apparent from the description and drawings, and from the claims. All publications, patents and patent applications cited herein are hereby expressly incorporated by reference for all purposes.

The invention provides methods for simultaneously identifying individual proteins in complex mixtures of biological molecules and quantifying the expression levels of those proteins, e.g., proteome analyses. The methods compare two or more samples of proteins, one of which can be considered as the standard sample and all others can be considered as samples under investigation. The proteins in the standard and investigated samples are subjected separately to a series of chemical modifications, i.e., differential chemical labeling, and fragmentation, e.g., by proteolytic digestion and/or other enzymatic reactions or physical fragmenting methodologies. The chemical modifications can be done before, or after, or before and after fragmentation/ digestion of the polypeptide into peptides.

Peptides derived from the standard and the investigated samples are labeled with chemical residues of different mass, but of similar properties, such that peptides with the same sequence from both samples are eluted together in the separation procedure and their ionization and detection properties regarding the mass spectrometry are very similar. Differential chemical labeling can be performed on reactive functional groups on some or all of the carboxy- and/or amino- termini of proteins and peptides and/or on selected amino acid side chains. A combination of chemical labeling, proteolytic digestion and other enzymatic reaction steps, physical fragmentation and/or fractionation can provide access to a variety of residues to general different specifically labeled peptides to enhance the overall selectivity of the procedure.

The standard and the investigated samples are combined, subjected to multidimensional chromatographic separation, and analyzed by mass spectrometry methods. Mass spectrometry data is processed by special software, which allows for identification and quantification of peptides and proteins.

Depending on the complexity and composition of the protein samples, it may be desirable, or be necessary, to perform protein fractionation using such methods as size exclusion, ion exchange, reverse phase, or other methods of affinity purifications prior to one or more chemical modification steps, proteolytic digestion or other enzymatic reaction steps, or physical fragmentation steps.

The combined mixtures of peptides are first separated by a chromatography method, such as a multidimensional liquid chromatography, system, before being fed into a coupled mass spectrometry device, such as a tandem mass spectrometry device. The combination of multidimensional liquid chromatography and tandem mass spectrometry can be called "LC-LC-MS/MS." LC-LC-MS/MS was first developed by Link A. and Yates J. R., as described, e.g., by Link (1999) Nature Biotechnology 17:676-682; Link (1999) Electrophoresis 18:1314-1334; Washburn, MP; Wolters, D; Yates, JR , Nature Biotechnology 2001 Mar, 19(3):242-7.

In practicing the methods of the invention, proteins can be first substantially or partially isolated from the biological samples of interest. The polypeptides can be treated before selective differential labeling; for example, they can be denatured, reduced, preparations can be desalted, and the like. Conversion of samples of proteins into mixtures of differentially labeled peptides can include preliminary chemical and/or enzymatic modification of side groups and/or termini; proteolytic digestion or fragmentation; post-digestion or post-fragmentation chemical and/or enzymatic modification of side groups and/or termini.

The differentially modified polypeptides and peptides are then combined into one or more peptide mixtures. Solvent or other reagents can be removed, neutralized or diluted, if desired or necessary. The buffer can be modified, or, the peptides can be redissolved in one or more different buffers, such as a "MudPIT" (see below) loading buffer. The peptide mixture is then loaded onto chromatography column, such as a liquid chromatography column, a 2D capillary column or a multidimensional chromatography column, to generate an eluate.

The eluate is fed into a mass spectrometer, such as a tandem mass spectrometer. In one aspect, an LC ESI MS and MS/MS analysis is complete. Finally, data output is processed by appropriate software using database searching and data analysis.

In practicing the methods of the invention, high yields of peptides can generated for mass spectrograph analysis. Two or more samples can be differentially labeled by selective labeling of each sample. Peptide modifications, i.e., labeling, are stable. Reagents having differing masses or reactive groups can be chosen to maximize the number of reactive groups and differentially labeled samples, thus allowing for a multiplex analysis of sample, polypeptides and peptides. In one aspect, a "MudPIT" protocol is used for peptide analysis, as described herein. The methods of the invention can be fully automated and can essentially analyze every protein in a sample.

Definitions

Unless defined otherwise, all technical and scientific terms used herein have the meaning commonly understood by a person skilled in the art to which this invention belongs. As used herein, the following terms have the meanings ascribed to them unless specified otherwise.

As used herein, the term "alkyl" is used to refer to a genus of compounds including branched or unbranched, saturated or unsaturated, monovalent hydrocarbon radicals, including substituted derivatives and equivalents thereof. In one aspect, the hydrocarbons have from about 1 to about 100 carbons, about 1 to about 50 carbons or about 1 to about 30 carbons, about 1 to about 20 carbons, about 1 to about 10 carbons. When the alkyl group has from about 1 to 6 carbon atoms, it is referred to as a "lower alkyl." Suitable alkyl radicals include, e.g., structures containing one or more methylene, methine and/or methyne groups arranged in acyclic and/or cyclic forms. Branched structures have a branching motif similar to isopropyl, tert-butyl isobutyl, 2-ethylpropyl, etc. As used herein, the term encompasses "substituted alkyls." "Substituted alkyl" refers to alkyl as just described including one or more functional groups such as lower alkyl, aryl, acyl, halogen (i.e., alkylhalos, e.g., CF3), hydroxy, amino, alkoxy, alkylamino, acylamino, thioamido, acyloxy, aryloxy, arylamino, aryloxyalkyl, mercapto, thia, aza, oxo, both saturated and unsaturated cyclic hydrocarbons, heterocycles and the like. These groups may be attached to any carbon of the alkyl moiety. Additionally, these groups may be pendent from, or integral to, the alkyl chain.

The term "alkoxy" is used herein to refer to the to a COR group, where R is a lower alkyl, substituted lower alkyl, aryl, substituted aryl, arylalkyl or substituted arylalkyl wherein the alkyl, aryl, substituted aryl, arylalkyl and substituted arylalkyl groups are as described herein. Suitable alkoxy radicals include, for example, methoxy, ethoxy, phenoxy, substituted phenoxy, benzyloxy phenethyloxy, tert.-butoxy, etc. The term "aryl" is used herein to refer to an aromatic substituent that may be a single aromatic ring or multiple aromatic rings which are fused together, linked covalently, or linked to a common group such as a methylene or ethylene moiety. The common linking group may also be a carbonyl as in benzophenone. The aromatic ring(s) may include phenyl, naphthyl, biphenyl, diphenylmethyl and benzophenone among others. The term "aryl" encompasses "arylalkyl." "Substituted aryl" refers to aryl as just described including one or more functional groups such as lower alkyl, acyl, halogen, alkylhalos (e.g., CF3), hydroxy, amino, alkoxy, alkylamino, acylamino, acyloxy, phenoxy, mercapto and both saturated and unsaturated cyclic hydrocarbons which are fused to the aromatic ring(s), linked covalently or linked to a common group such as a methylene or ethylene moiety. The linking group may also be a carbonyl such as in cyclohexyl phenyl ketone. The term "substituted aryl" encompasses "substituted arylalkyl."

The term "arylalkyl" is used herein to refer to a subset of "aryl" in which the aryl group is further attached to an alkyl group, as defined herein.

The term "biotin" as used herein refers to any natural or synthetic biotin or variant thereof, which are well known in the art; ligands for biotin, and ways to modify the affinity of biotin for a ligand, are also well known in the art; see, e.g., U.S. Patent Nos. 6,242,610; 6,150,123; 6,096,508; 6,083,712; 6,022,688; 5,998,155; 5,487,975.

The phrase "labeling reagents which ... do not differ in ionization and detection properties in mass spectrographic analysis" means that the amount and/or mass sequence of the labeling reagents can be detected using the same mass spectrographic conditions and detection devices.

The term "polypeptide" includes natural and synthetic polypeptides, or mimetics, which can be either entirely composed of synthetic, non-natural analogues of amino acids, or, they can be chimeric molecules of partly natural peptide amino acids and partly non-natural analogs of amino acids. The term "polypeptide" as used herein includes proteins and peptides of all sizes.

The term "sample" as used herein includes any polypeptide-containing sample, including samples from natural sources, or, entirely synthetic samples.

The term "column" as used herein means any substrate surface, including beads, filaments, arrays, tubes and the like.

The phrase "do not differ in chromatographic retention properties" as used herein means that two compositions have substantially, but not necessary exactly, the same retention properties in a chromatograph, such as a liquid chromatograph. For example, two compositions do not differ in chromatographic retention properties if they elute together, i.e., they elute in what a skilled artisan would consider the same elution fraction.

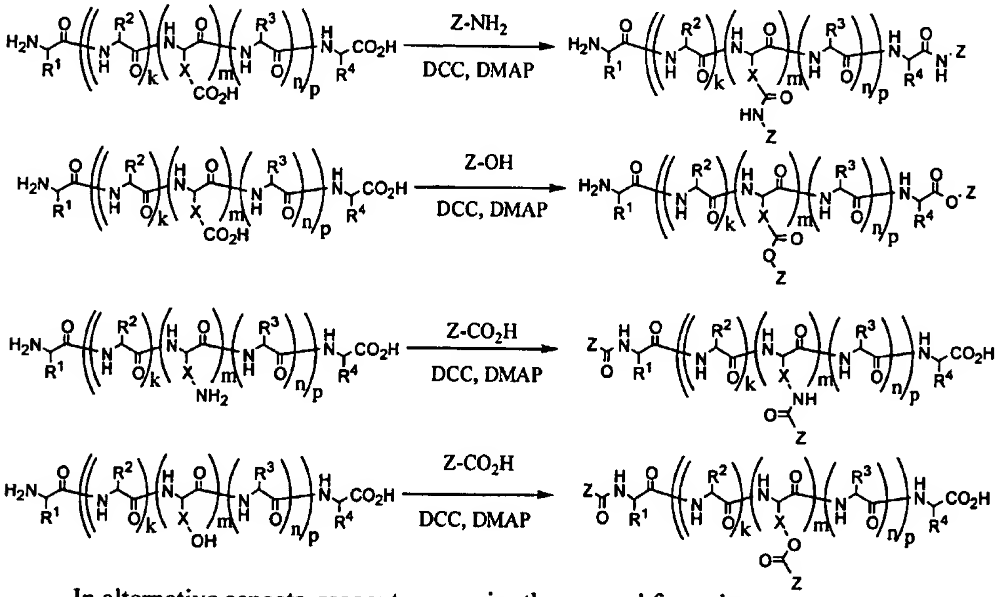
Differential labeling of peptides and polypeptides

In practicing the methods of the invention, proteins and peptides are subjected to a series of chemical modifications, i.e., differential chemical labeling. The chemical modifications can be done before, or after, or before and after fragmentation/ digestion of the polypeptide into peptides. Differential labeling reagents can differ in their isotope composition (i.e., isotopical reagents), in their structural composition (i.e., homologous reagents), but by a rather small fragment which change does not alter the properties stated above, i.e., the labeling reagent differ in molecular mass but do not differ in chromatographic retention properties and do not differ in ionization and detection properties in mass spectrographic analysis, and the differences in molecular mass are distinguishable by mass spectrographic analysis.

In one aspect of the invention, mixtures of polypeptides and/or peptides coming from the "standard" protein sample and the "investigated" protein sample(s) are labeled separately with differential reagents, or, one sample is labeled and other sample remains unlabeled. As noted above, these differential reagents differ in molecular mass, but do not differ in retention properties regarding the separation method used (e.g., chromatography) and the mass spectrometry methods used will not detect different ionization and detection properties. Thus, these differential reagents differ either in their isotope composition (i.e., they are isotopical reagents) or they differ structurally by a rather small fragment which change does not alter the properties stated above (i.e., they are homologous reagents).

Differential chemical labeling can include esterification of C-termini, amidation of C-termini and/or acylation of N-termini. Esterification targets C-termini of peptides and carboxylic acid groups in amino acid side chains. Amidation targets C-termini of peptides and carboxylic acid groups in amino acid side chains. Amidation may require protection of amine groups first. Acylation targets N-termini of peptides and amino and hydroxy groups in amino acid side chains. Acylation may require protection of carboxylic groups first.

The skilled artisan will recognize that the chemical syntheses and differential chemical labeling of peptides and polypeptides (e.g., esterification, amidation, and acylation) used to practice the methods of the invention can be by a variety of procedures and methodologies, which are well described in the scientific and patent literature, e.g., Organic Syntheses Collective Volumes, Gilman et al. (Eds), John Wiley & Sons, Inc., NY; Venuti (1989) Pharm. Res. 6: 867-873; the Beilstein Handbook of Organic Chemistry (Beilstein Institut fuer Literatur der Organischen Chemie, Frankfurt, Germany); Beilstein online database and references obtainable therein; "Organic Chemistry," Morrison & Boyd, 7th edition, 1999, Prentice-Hall, Upper Saddle River, NJ. The invention can be practiced in conjunction with any method or protocol known in the art, which are well described in the scientific and patent literature. For example, the esterification, amidation, and acylation reactions may be performed on the mixtures of peptides in a fashion similar to other reaction of these types already described in prior art, such as:



In alternative aspects, reagents comprise the general formulae:

i. Z^AOH and Z^BOH to esterify peptide C-terminals and/or Glu and Asp side chains;

ii. Z^ANH_2 / Z^BNH_2 to form amide bond with peptide C-terminals and/or Glu and Asp side chains; or

iii. Z^ACO_2H / Z^BCO_2H to form amide bond with peptide N-terminals and/or Lys and Arg side chains;

wherein Z^A and Z^B independently of one another can be R-Z^1-A^1-Z^2-A^2-Z^3-A^3-Z^4-A^4- , and Z^1, Z^2, Z^3, and Z^4 independently of one another can be selected from O, OC(O), OC(S), OC(O)O, OC(O)NR, OC(S)NR, OSiRR^1, S, SC(O), SC(S), SS, S(O), S(O$_2$), NR, NRR^{1+}, C(O), C(O)O, C(S), C(S)O, C(O)S, C(O)NR, C(S)NR, SiRR^1, (Si(RR^1)O)n, SnRR^1, Sn(RR^1)O, BR(OR^1), BRR^1, B(OR)(OR^1) , OBR(OR^1), OBRR^1, OB(OR)(OR^1), or, Z^1, Z^2, Z^3, and Z^4 independently of one another may be absent, and R is an alkyl group; and, A^1, A^2, A^3, and A^4 independently of one another can be selected from (CRR^1)n, and R is an alkyl group. In alternative aspects, some single C-C bonds from (CRR^1)n may be replaced with double or triple bonds, in which case some groups R and R^1 will be absent, (CRR^1)n can be an *o*-arylene, an *m*-arylene, or a *p*-arylene with up to 6 substituents, carbocyclic, bicyclic, or tricyclic fragments with up to 8 atoms in the cycle with or without heteroatoms (O, N, S) and with or without substituents, or A^1, A^2, A^3, and A^4 independently of one another can be absent; R, R^1, independently from other R and R^1 in

Z^1 - Z^4 and independently from other R and R^1 in A^1 - A^4, can be hydrogen, halogen or an alkyl group, such as an alkenyl, an alkynyl or an aryl group; n in Z^1 - Z^4, independent of n in A^1 - A^4, is an integer that can have value from 0 to about 51; 0 to about 41; 0 to about 31; 0 to about 21, 0 to about 11; 0 to about 6;

In alternative aspects, Z^A has the same structure as Z^B, but they have different isotope compositions. Any isotope may be used. In alternative aspects, if Z^A contains *x* number of protons, Z^B may contain *y* number of deuterons in the place of protons, and, correspondingly, *x* - *y* number of protons remaining; and/or if Z^A contains *x* number of borons-10, Z^B may contain *y* number of borons-11 in the place of borons-10, and, correspondingly, *x* - *y* number of borons-10 remaining; and/or if Z^A contains *x* number of carbons-12, Z^B may contain *y* number of carbons-13 in the place of carbons-12, and, correspondingly, *x* - *y* number of carbons-12 remaining; and/or if Z^A contains *x* number of nitrogens-14, Z^B may contain *y* number of nitrogens-15 in the place of nitrogens-14, and, correspondingly, *x* - *y* number of nitrogens-14 remaining; and/or if Z^A contains *x* number of sulfurs-32, Z^B may contain *y* number of sulfurs-34 in the place of sulfurs-32, and, correspondingly, *x* - *y* number of sulfurs-32 remaining; and so on for all elements which may be present and have different stable isotopes; *x* and *y* are whole numbers such that *x* is greater than *y*. In one aspect, *x* and *y* are between 1 and about 11, between 1 and about 21, between 1 and about 31, between 1 and about 41, between 1 and about 51.

In alternative aspects, reagent pairs/series comprise the general formulae:

i. $CD_3(CD_2)_nOH$ / $CH_3(CH_2)_nOH$ to esterify peptide C-terminals, where n = 0, 1, 2, …, y; (delta mass = 3 + 2n);

ii. $CD_3(CD_2)_nNH_2$ / $CH_3(CH_2)_nNH_2$ to form amide bond with peptide C-terminals where n = 0, 1, 2, …, y (delta mass = 3+ 2n);

iii. $D(CD_2)_nCO_2H$ / $H(CH_2)_nCO_2H$ to form amide bond with peptide N-terminals, where n = 0, 1, 2, …, y (delta mass = 1+2n);

wherein y is an integer that can have value of about 51; about 41; about 31; about 21, about 11; about 6, or between about 5 and 51.

Other exemplary reagents can be presented by general formulae:

i. Z^AOH and Z^BOH to esterify peptide C-terminals;

ii. Z^ANH_2 / Z^BNH_2 to form an amide bond with peptide C-terminals;

iii. Z^ACO_2H / Z^BCO_2H to form an amide bond with peptide N-terminals;

wherein Z^A and Z^B can be $R\text{-}Z^1\text{-}A^1\text{-}Z^2\text{-}A^2\text{-}Z^3\text{-}A^3\text{-}Z^4\text{-}A^4\text{-}$

and Z^1, Z^2, Z^3, and Z^4, independently of one another, can be selected from O, OC(O), OC(S), OC(O)O, OC(O)NR, OC(S)NR, $OSiRR^1$, S, SC(O), SC(S), SS, S(O), $S(O_2)$, NR, NRR^{1+}, C(O), C(O)O, C(S), C(S)O, C(O)S, C(O)NR, C(S)NR, $SiRR^1$, $(Si(RR^1)O)n$, $SnRR^1$, $Sn(RR^1)O$, $BR(OR^1)$, BRR^1, $B(OR)(OR^1)$, $OBR(OR^1)$, $OBRR^1$, or $OB(OR)(OR^1)$; or, Z^1, Z^2, Z^3, and Z^4, independently of one another, can be absent, and, R is an alkyl group;

A^1, A^2, A^3, and A^4, independently of one another, can be a moiety comprising the general formulae $(CRR^1)n$. In alternative aspects, single C-C bonds in some $(CRR^1)n$ groups may be replaced with double or triple bonds, in which case some groups R and R^1 will be absent, or $(CRR^1)n$ can be an *o*-arylene, an *m*-arylene, or a *p*-arylene with up to 6 substituents, or a carbocyclic, a bicyclic, or a tricyclic fragments with up to 8 atoms in the cycle, with or without heteroatoms (e.g., O, N or S atoms), or, with or without substituents, or, A^1 - A^4 independently of one another may be absent;

In alternative aspects, R, R^1, independently from other R and R^1 in Z^1 - Z^4 and independently from other R and R^1 in A^1 - A^4, can be a hydrogen atom, a halogen or an alkyl group, such as an alkenyl, an alkynyl or an aryl group;

In alternative aspects, n in Z^1 - Z^4 is independent of n in A^1 - A^4 and is an integer that can have value of about 51; about 41; about 31; about 21, about 11; about 6.

In alternative aspects, Z^A has a similar structure to that of Z^B, but Z^A has *x* extra $-CH_2-$ fragment(s) in one or more A^1 - A^4 fragments, and/or Z^A has *x* extra $-CF_2-$ fragment(s) in one or more A^1 - A^4 fragments. Alternatively, Z^A can contain *x* number of protons and Z^B may contain *y* number of halogens in the place of protons. Alternatively, where Z^A contains *x* number of protons and Z^B contains *y* number of halogens, there are *x* - *y* number of protons remaining in one or more A^1 - A^4 fragments; and/or Z^A has *x* extra –O- fragment(s) in one or more A^1 - A^4 fragments; and/or Z^A has *x* extra –S- fragment(s) in one or more A^1 - A^4 fragments; and/or if Z^A contains *x* number of –O- fragment(s), Z^B may contain *y* number of –S- fragment(s) in the place of –O- fragment(s), and, correspondingly,

x - *y* number of –O- fragment(s) remaining in one or more A^1 - A^4 fragments; and the like.

In alternative aspects, *x* and *y* are integers that can have value of between 1 about 51; of between 1 about 41; of between 1 about 31; of between 1 about 21, of between 1 about 11; of between 1 about 6, such that *x* is greater than *y*.

Exemplary homologous reagents pairs/series are

i. $CH_3(CH_2)_nOH/CH_3(CH_2)_{n+m}OH$ to esterify peptide C-terminals, where n = 0, 1, 2, ..., y; m = 1, 2, ..., y (delta mass = 14m)

ii. $CH_3(CH_2)_n NH_2$ / $CH_3(CH_2)_{n+m}NH_2$ to form amide bond with peptide C-terminals, where n = 0, 1, 2, ..., y; m = 1, 2, ..., y (delta mass = 14m)

iii. $H(CH_2)_nCO_2H$ / $H(CH_2)_{n+m}CO_2H$ to form amide bond with peptide N-terminals, where n = 0, 1, 2, ..., y; m = 1, 2, ..., y (delta mass = 14m)

wherein y is an integer that can have value of about 51; about 41; about 31; about 21, about 11; about 6, or between about 5 and 51.

Methods for peptide/protein separation and detection

The methods of the invention use chromatographic techniques to separate tagged polypeptides and peptides. In one aspect, a liquid chromatography is used, e.g., a multidimensional liquid chromatography. The chromatogram eluate is coupled to a mass spectrometer, such as a tandem mass spectrometry device (e.g., a "LC-LC-MS/MS" system). Any variation and equivalent thereof can be used to separate and detect peptides. LC-LC-MS/MS was first developed by Link A. and Yates J. R., as described, e.g., in (Link (1999) Nature Biotechnology 17:676-682; Link (2000) Electrophoresis 18, 1314-1334. In one aspect, the LC-LC-MS/MS technique is used; it is effective for complexed peptide separation and it is easily automated. LC-LC-MS/MS is commonly known by the acronym "MudPIT," for "Multi-dimensional Protein Identification Technique."

Variations and equivalents of LC-LC-MS/MS used in the methods of the invention include methodologies involving reversed phase columns coupled to either cation exchange columns (as described, e.g., by Opiteck (1997) Anal. Chem. 69:1518-1524; or, size exclusion columns (as described, e.g., by Opiteck (1997) Anal. Biochem. 258:349-361). In one aspect, an LC-LC-MS/MS technique uses a mixed bed microcapillary column containing strong cation exchange (SCX) and reversed phase (RPC) resins. Other exemplary alternatives include protein fractionation combined with one-dimensional LC-ESI MS/MS or peptide fractionation combined MALDI MS/MS.

Depending on the complexity or the property of the protein samples, any protein fractionation method, including size exclusion chromatography, ion exchange chromatography, reverse phase chromatography, or any of the possible affinity purifications, can be introduced prior to labeling and proteolysis. In some circumstances, use of several different methods may be necessary to identify all proteins or specific proteins in a sample.

Sequence analysis and quantification

Both quantity and sequence identity of the protein from which the modified peptide originated can be determined by a mass spectrometry device, such as a "multistage mass spectrometry" (MS). This can be achieved by the operation of the mass spectrometer in a dual mode in which it alternates in successive scans between measuring the relative quantities of peptides eluting from the capillary column and recording the sequence information of selected peptides. Peptides are quantified by measuring in the MS mode the relative signal intensities for pairs or series of peptide ions of identical sequence that are tagged differentially, which therefore differ in mass by the mass differential encoded within the differential labeling reagents.

Peptide sequence information can be automatically generated by selecting peptide ions of a particular mass-to-charge (m/z) ratio for collision-induced dissociation (CID) in the mass spectrometer operating in the tandem MS mode, as described, e.g., by Link (1997) Electrophoresis 18:1314-1334; Gygi (1999) Nature Biotechnol. 17:994-999; Gygi (1999) Cell Biol. 19:1720-1730.

The resulting tandem mass spectra can be correlated to sequence databases to identify the protein from which the sequenced peptide originated. Exemplary commercial available softwares include TURBO SEQUEST™ by Thermo Finnigan, San Jose, CA; MASSSCOT™ by Matrix Science, SONAR MS/MS™ by Proteometrics. Routine software modifications may be necessary for automated relative quantification.

Mass spectrometry devices

In the methods of the invention use mass spectrometry to identify and quantify differentially labeled peptides and polypeptides. Any mass spectrometry system can be used. In one aspect of the invention, combined mixtures of peptides are separated by a chromatography method comprising multidimensional liquid chromatography coupled to tandem mass spectrometry, or, "LC-LC-MS/MS," see, e.g., Link (1999) Biotechnology 17:676-682; Link (1999) Electrophoresis 18:1314-1334. Exemplary, mass spectrometry devices include those incorporating matrix-assisted laser desorption-ionization-time-of-flight (MALDI-TOF) mass spectrometry (see, e.g., Isola (2001) Anal. Chem. 73:2126-2131; Van de Water (2000) Methods Mol. Biol. 146:453-459; Griffin (2000) Trends Biotechnol. 18:77-84; Ross (2000) Biotechniques 29:620-626, 628-629). The inherent high molecular weight resolution of MALDI-TOF MS conveys high specificity and good signal-to-noise ratio for performing accurate quantitation.

Use of mass spectrometry, including MALDI-TOF MS, and its use in detecting nucleic acid hybridization and in nucleic acid sequencing, is well known in the art, see, e.g., U.S. Patent Nos. 6,258,538; 6,238,871; 6,238,869; 6,235,478; 6,232,066; 6,228,654; 6,225,450; 6,051,378; 6,043,031.

Fragmentation and proteolytic digestion

In practicing the methods of the invention, polypeptides are fragmented, e.g., by proteolytic, i.e., enzymatic, digestion and/or other enzymatic reactions or physical fragmenting methodologies. The fragmentation can be done before and/or after reacting the peptides/ polypeptides with the labeling reagents used in the methods of the invention.

Methods for proteolytic cleavage of polypeptides are well known in the art, e.g., enzymes include trypsin (see, e.g., U.S. Patent No. 6,177,268; 4,973,554), chymotrypsin (see, e.g., U.S. Patent No. 4,695,458; 5,252,463), elastase (see, e.g., U.S. Patent No. 4,071,410); subtilisin (see, e.g., U.S. Patent No. 5,837,516) and the like.

In one aspect, a chimeric labeling reagent of the invention includes a cleavable linker. Exemplary cleavable linker sequences include, e.g., Factor Xa or enterokinase (Invitrogen, San Diego CA). Other purification facilitating domains can be used, such as metal chelating peptides, e.g., polyhistidine tracts and histidine-tryptophan modules that allow purification on immobilized metals, protein A domains that allow purification on immobilized immunoglobulin, and the domain utilized in the FLAGS extension/affinity purification system (Immunex Corp, Seattle WA).

Biological Samples

The methods are based on comparison of two or more samples of proteins, one of which can be considered as the standard sample and all others can be considered as samples under investigation. For example, in one aspect, the invention provides a method for quantifying changes in protein expression between at least two cellular states, such as, an activated cell versus a resting cell, a normal cell versus a cancerous cell, a stem cell versus a differentiated cell, an injured cell or infected cell versus an uninjured cell or uninfected cell; or, for defining the expressed proteins associated with a given cellular state.

Sample can be derived from any biological source, including cells from, e.g., bacteria, insects, yeast, mammals and the like. Cells can be harvested from any body fluid or tissue source, or, they can be *in vitro* cell lines or cell cultures.

Detection Devices and Methods

The devices and methods of the invention can also incorporate in whole or in part designs of detection devices as described, e.g., in U.S. Patent Nos. 6,197,503; 6,197,498; 6,150,147; 6,083,763; 6,066,448; 6,045,996; 6,025,601; 5,599,695; 5,981,956; 5,698,089; 5,578,832; 5,632,957.

A number of embodiments of the invention have been described. Nevertheless, it will be understood that various modifications may be made without departing from the spirit and scope of the invention.

References

Unless otherwise indicated, all references cited herein (supra and infra) are incorporated by reference in their entirety.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R.: Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 17(10):994-9 (Oct) 1999.

Hopkins MJ, Sharp R, Macfarlane GT.: Age and disease related changes in intestinal bacterial populations assessed by cell culture, 16S rRNA abundance, and community cellular fatty acid profiles. *Gut* 48(2):198-205 (Feb) 2001.

Ritchie NJ, Schutter ME, Dick RP, Myrold DD.: Use of length heterogeneity PCR and fatty acid methyl ester profiles to characterize microbial communities in soil.*Appl Environ Microbiol* 66(4):1668-75 (Apr) 2000.

Khan AA, Wang RF, Cao WW, Franklin W, Cerniglia CE.: Reclassification of a polycyclic aromatic hydrocarbon-metabolizing bacterium, Beijerinckia sp. strain B1, as Sphingomonas yanoikuyae by fatty acid analysis, protein pattern analysis, DNA-DNA hybridization, and 16S ribosomal DNA sequencing. *Int J Syst Bacteriol* 46(2):466-9 (Apr) 1996.

Peltroche-Llacsahuanga H, Schmidt S, Lutticken R, Haase G.: Discriminative power of fatty acid methyl ester (FAME) analysis using the microbial identification system (MIS) for Candida (Torulopsis) glabrata and Saccharomyces cerevisiae. *Diagn Microbiol Infect Dis* 38(4):213-21 (Dec) 2000.

SA Gerber et al.: Analysis of rates of multiple enzymes in cell lysates by electrospray ionization mass spectrometry. *J. Am. Chem. Soc.* 121:1102-3 1999.

www.genomeweb.com
David Goodlett discusses the latest in genomics – ICAT reagents
Written by: Marian Moser Jones
Dec 20, 2000

WO0011208; Filed Aug 25, 1999, Published March 2, 2000. Aebersold RH, Gelb MH, Gygi, SP, Scott CR, Turecek F, Gerber SA, Rist B: Rapid quantitative analysis of proteins or protein function in complex mixtures.

WO9905221; Filed July 27 1998, Published Feb. 4,1999. Cummins WJ, West RM, Smith JA: Cyanine Dyes.

US4876350; Filed Dec 16, 1987, Issued Oct 24, 1989. McGarrity J, Tenud L: Process for the production of (+) biotin.

US5776723; Filed Feb 8, 1996, Issued July 7, 1998. Herold CD, O'Hagan M: Rapid detection of mycobacterium tuberculosis.

US6136173; Filed June 24, 1996, Issued Oct. 24, 2000. Anderson NL, Anderson NG, Goodman J: Automated system for two-dimensional electrophoresis.

US6127134; Filed April 20, 1995, Issued Oct. 3, 2000. Minden J, Waggoner A: Difference gel electrophoresis using matched multiple dyes.

US6064754; Filed Dec 1, 1997, Issued May 16, 2000. Parekh RB, Amess R, Bruce JA, Prime SB, Platt AE, Stoney RM: Computer-assisted methods and apparatus for identification and characterization of biomolecules in a biological sample.

US6013165; Filed May 22, 1998, Issued Jan 11,2000. Wiktorowicz JE, Raysberg Y: Electrophoresis apparatus and method.

Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K Editors.Current Protocols In Molecular Biology, Vol 2. John Wiley & Sons, Inc, © 2001, 10.21.4-10.21.6, 10.22.5-10.22.10, 10.22.14, 10.22.15-10.22.20.

Sambrook J, Russell DW Editors. Molecular Cloning A Laboratory Manual 3rd ed. Cold Spring Harbor Laboratory Press, New York, © 2001, 18.3, 18.62, 18.66.

Alting-Mecs MA and Short JM: Polycos vectors: a system for packaging filamentous phage and phagemid vectors using lambda phage packaging extracts. *Gene* 137:1, 93-100, 1993.

Arkin AP and Youvan DC: An algorithm for protein engineering: simulations of recursive ensemble mutagenesis. *Proc Natl Acad Sci USA* 89(16):7811-7815, (Aug 15) 1992.

Arnold FH: Protein engineering for unusual environments. *Current Opinion in Biotechnology* 4(4):450-455, 1993.

Ausubel FM, et al Editors. Current Protocols in Molecular Biology, Vols. 1 and 2 and supplements. (a.k.a. "The Red Book") Greene Publishing Assoc., Brooklyn, NY, ©1987.

Ausubel FM, et al Editors. Current Protocols in Molecular Biology, Vols. 1 and 2 and supplements. (a.k.a. "The Red Book") Greene Publishing Assoc., Brooklyn, NY, ©1989.

Ausubel FM, et al Editors. Short Protocols in Molecular Biology: A Compendium of Methods from Current Protocols in Molecular Biology. Greene Publishing Assoc., Brooklyn, NY, ©1989.

Ausubel FM, et al Editors. Short Protocols in Molecular Biology: A Compendium of Methods from Current Protocols in Molecular Biology, 2nd Edition. Greene Publishing Assoc., Brooklyn, NY, ©1992.

Barbas CF 3d, Bain JD, Hoekstra DM, Lerner RA: Semisynthetic combinatorial antibody libraries: a chemical solution to the diversity problem. *Proc Natl Acad Sci USA* 89(10):4457-4461, 1992.

Bardwell AJ, Bardwell L, Johnson DK, Friedberg EC: Yeast DNA recombination and repair proteins Rad1 and Rad10 constitute a complex in vivo mediated by localized hydrophobic domains. *Mol Microbiol* 8(6):1177-1188, 1993.

Barret AJ, et al., eds.: Enzyme Nomenclature: Recommendations of the Nomenclature

Committee of the International Union of Biochemistry and Molecular Biology. San Diego: Academic Press, Inc., 1992.
Bartel P, Chien CT, Sternglanz R, Fields S: Elimination of false positives that arise in using the two-hybrid system. *Biotechniques* 14(6):920-924, 1993.
Beaudry AA and Joyce GF: Directed evolution of an RNA enzyme. *Science* 257(5070):635-641, 1992.
Berger and Kimmel, Methods in Enzymology, Volume 152, Guide to Molecular Cloning Techniques. Academic Press, Inc., San Diego, CA, ©1987. (Cumulative Subject Index: Volumes 135-139, 141-167, 1990, 272 pp.)
Bevan M: Binary Agrobacterium vectors for plant transformation. *Nucleic Acids Research* 12(22):8711-21, 1984.
Biocca S, Pierandrei-Amaldi P, Cattaneo A: Intracellular expression of anti-p21ras single chain Fv fragments inhibits meiotic maturation of xenopus oocytes. *Biochem Biophys Res Commun* 197(2):422-427, 1993.
Bird et al. *Plant Mol Biol* 11:651, 1988..
Bogerd HP, Fridell RA, Blair WS, Cullen BR: Genetic evidence that the Tat proteins of human immunodeficiency virus types 1 and 2 can multimerize in the eukaryotic cell nucleus. *J Virol* 67(8):5030-5034, 1993.
Boyce COL, ed.: Novo's Handbook of Practical Biotechnology. 2nd ed. Bagsvaerd, Denmark, 1986.
Brederode FT, Koper-Zawrthoff EC, Bol JF: Complete nucleotide sequence of alfalfa mosaic virus RNA 4. *Nucleic Acids Research* 8(10):2213-23, 1980.
Breitling F, Dubel S, Seehaus T, Klewinghaus I, Little M: A surface expression vector for antibody screening. *Gene* 104(2):147-153, 1991.
Brown NL, Smith M: Cleavage specificity of the restriction endonuclease isolated from Haemophilus gallinarum (Hga I). *Proc Natl Acad Sci U S A* 74(8):3213-6, (Aug) 1977.
Burton DR, Barbas CF 3d, Persson MA, Koenig S, Chanock RM, Lerner RA: A large array of human monoclonal antibodies to type 1 human immunodeficiency virus from combinatorial libraries of asymptomatic seropositive individuals. *Proc Natl Acad Sci U S A* 88(22):10134-7, (Nov 15) 1991.
Caldwell RC and Joyce GF: Randomization of genes by PCR mutagenesis. *PCR Methods Appl* 2(10):28-33, 1992.
Caton AJ and Koprowski H: Influenze virus hemagglutinin-specific antibodies isolatedf froma combinatorial expression library are closely related to the immune response of the donor. *Proc Natl Acad Sci USA* 87(16):6450-6454, 1990.
Chakraborty T, Martin JF, Olson EN: Analysis of the oligomerization of myogenin and E2A products in vivo using a two-hybrid assay system. *J Biol Chem* 267(25):17498-501, 1992.
Chang CN, Landolfi NF, Queen C: Expression of antibody Fab domains on bacteriophage surfaces. Potential use for antibody selection. *J Immunol* 147(10):3610-4, (Nov 15) 1991.
Chaudhary VK, Batra JK, Gallo MG, Willingham MC, FitzGerald DJ, Pastan I: A rapid method of cloning functional variable-region antibody genes in Escherichia coli as single-chain immunotoxins. *Proc Natl Acad Sci USA* 87(3):1066-1070, 1990.
Chien CT, Bartel PL, Sternglanz R, Fields S: The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc Natl Acad Sci USA* 88(21):9578-9582, 1991.
Chiswell DJ, McCafferty J: Phage antibodies: will new 'coliclonal' antibodies replace monoclonal antibodies? *Trends Biotechnol* 10(3):80-84, 1992.

Chothia C and Lesk AM: Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* 196)4):901-917, 1987.

Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR, et al: Conformations of immunoglobulin hypervariable regions. *Nature* 342(6252):877-883, 1989.

Clackson T, Hoogenboom HR, Griffiths AD, Winter G: Making antibody fragments using phage display libraries. *Nature* 352(6336):624-628, 1991.

Conrad M, Topal MD: DNA and spermidine provide a switch mechanism to regulate the activity of restriction enzyme *Nae* I. *Proc Natl Acad Sci U S A* 86(24):9707-11, (Dec) 1989.

Coruzzi G, Broglie R, Edwards C, Chua NH: Tissue-specific and light-regulated expression of a pea nuclear gene encoding the small subunit of ribulose-1,5-bisphosphate carboxylase. *EMBO J* 3(8):1671-9, 1984.

Dasmahapatra B, DiDomenico B, Dwyer S, Ma J, Sadowski I, Schwartz J: A genetic system for studying the activity of a proteolytic enzyme. *Proc Natl Acad Sci USA* 89(9):4159-4162, 1992.

Davis LG, Dibner MD, Battey JF. Basic Methods in Molecular Biology. Elsevier, New York, NY, ©1986.

Delegrave S and Youvan DC. *Biotechnology Research* 11:1548-1552, 1993.

DeLong EF, Wu KY, Prezelin BB, Jovine RV: High abundance of Archaea in Antarctic marine picoplankton. *Nature* 371(6499):695-697, 1994.

Deng SJ, MacKenzie CR, Sadowska J, Michniewicz J, Young NM, Bundle Dr, Narang SA: Selection of antibody single-chain variable fragments with improved carbohydrate binding by phage display. *J Biol Chem* 269(13):9533-9538, 1994.

Drauz K, Waldman H, eds.: Enzyme Catalysis in Organic Synthesis: A Comprehensive Handbook. Vol. 1. New York: VCH Publishers, 1995.

Drauz K, Waldman H, eds.: Enzyme Catalysis in Organic Synthesis: A Comprehensive Handbook. Vol. 2. New York: VCH Publishers, 1995.

Duan L, Bagasra O, Laughlin MA, Oakes JW, Pomerantz RJ: Potent inhibition of human immunodeficiency virus type 1 replication by an intracellular anti-Rev single-chain antibody. *Proc Natl Acad Sci USA* 91(11):5075-5079, 1994.

Durfee T, Becherer K, Chen PL, Yeh SH, Yang Y, Kilburn AE, Lee WH, Elledge SJ: The retinoblastoma protein associates with the protein phosphatase type 1 catalytic subunit. *Genes Dev* 7(4):555-569, 1993.

Ellington AD and Szostak JW: In vitro selection of RNA molecules that bind specific ligands. *Nature* 346(6287):818-822, 1990.

Fields S and Song 0: A novel genetic system to detect protein-protein interactions. *Nature* 340(6230):245-246, 1989.

Firek S, Draper J, Owen MR, Gandecha A, Cockburn B, Whitelam GC: Secretion of a functional single-chain Fv protein in transgenic tobacco plants and cell suspension cultures. *Plant Mol Biol* 23(4):861-870, 1993.

Forsblom S, Rigler R, Ehrenberg M, Philipson L: Kinetic studies on the cleavage of adenovirus DNA by restriction endonuclease Eco RI. *Nucleic Acids Res* 3(12):3255-69, (Dec) 1976.

Foster GD, Taylor SC, eds.: Plant Virology Protocols: From Virus Isolation to Transgenic Resistance. Methods in Molecular Biology, Vol. 81. New Jersey: Humana Press Inc., 1998.

Franks F, ed.: Protein Biotechnology: Isolation, Characterization, and Stabilization. New Jersey: Humana Press Inc., 1993.

Germino FJ, Wang ZX, Weissman SM: Screening for in vivo protein-protein interactions. *Proc Natl Acad Sci USA* 90(3):933-937, 1993.

Gingeras TR, Brooks JE: Cloned restriction/modification system from Pseudomonas aeruginosa. *Proc Natl Acad Sci USA* 80(2):402-6, 1983 (Jan).

Gluzman Y: SV40-transformed simian cells support the replication of early SV40 mutants. *Cell* 23(1):175-182, 1981.

Godfrey T, West S, eds.: Industrial Enzymology. 2nd ed. London: Macmillan Press Ltd, 1996.

Gottschalk G: Bacterial Metabolism. 2nd ed. New York: Springer-Verlag Inc., 1986.

Gresshoff PM, ed.: Technology Transfer of Plant Biotechnology. Current Topics in Plant Molecular Biology. Boca Raton: CRC Press, 1997.

Griffin HG, Griffin AM, eds.: PCR Technology: Currrent Innovations. Boca Raton: CRC Press, Inc., 1994.

Gruber M, Schodin BA, Wilson ER, Kranz DM: Efficient tumor cell lysis mediated by a bispecific single chain antibody expressed in Escherichia coli. *J Immunol* 152(11):5368-5374, 1994.

Guarente L: Strategies for the identification of interacting proteins. *Proc Natl Acad Sci USA* 90(5):1639-1641, 1993.

Guilley H, Dudley RK, Jonard G, Balazs E, Richards KE: Transcription of Cauliflower mosaic virus DNA: detection of promoter sequences, and characterization of transcripts. *Cell* 30(3):763-73, 1982.

Hansen G, Chilton MD: Lessons in gene transfer to plants by a gifted microbe. *Curr Top Microbiol Immunol* 240:21-57, 1999.

Hardy CF, Sussel L, Shore D: A RAP1-interacting protein involved in transcriptional silencing and telomere length regulation. *Genes Dev* 6(5):801-814, 1992.

Hartmann HT, et al.: Plant Propagation: Principles and Practices. 6th ed. New Jersey: Prentice Hall, Inc., 1997.

Hawkins RE and Winter G: Cell selection strategies for making antibodies from variable gene libraries: trapping the memory pool. *Eur J Immunol* 22(3):867-870, 1992.

Holvoet P, Laroche Y, Lijnen HR, Van Hoef B, Brouwers E, De Cock F, Lauwereys M, Gansemans Y, Collen D: Biochemical characterization of single-chain chimeric plasminogen activators consisting of a single-chain Fv fragment of a fibrin-specific antibody and single-chain urokinase. *Eur J Biochem* 210(3):945-952, 1992.

Honjo T, Alt FW, Rabbitts TH (eds): *Immunoglobulin genes*. Academic Press: San Diego, CA, pp. 361-368, ©1989.

Hoogenboom HR, Griffiths AD, Johnson KS, Chiswell DJ, Judson P, Winter G: Multi-subunit proteins on the surface of filamentous phage: methodologies for displaying antibody (Fab) heavy and light chains. *Nucleic Acids Res* 19(15):4133-4137, 1991.

Huse WD, Sastry L, Iverson SA, Kang AS, Alting-Mees M, Burton DR, Benkovic SJ, Lerner RA: Generation of a large combinatorial library of the immunoglobulin repertoire in phage lambda. *Science* 246(4935):1275-1281, 1989.

Huston JS, Levinson D, Mudgett-Hunter M, Tai MS, Novotney J, Margolies MN, Ridge RJ, Bruccoleri RE, Haber E, Crea R, et al: Protein engineering of antibody binding sites: recovery of specific activity in an anti-digoxin single-chain Fv analogue produced in Escherichia coli. *Proc Natl Acad Sci USA* 85(16):5879-5883, 1988.

Ivan Lefkovits, Editor. Immunology methods manual : the comprehensive sourcebook of techniques. Academic Press, San Diego, ©1997.

Iwabuchi K, Li B, Bartel P, Fields S: Use of the two-hybrid system to identify the domain of p53 involved in oligomerization. *Oncogene* 8(6):1693-1696, 1993.

Jackson AL, Pahl PM, Harrison K, Rosamond J, Sclafani RA: Cell cycle regulation of the yeast Cdc7 protein kinase by association with the Dbf4 protein. *Mol Cell Biol* 13(5):2899-2908, 1993.

Johnson S and Bird RE: *Methods Enzymol* 203:88, 1991.

Kabat et al: Sequences of Proteins of Immunological Interest, 4th Ed. U.S. Department of Health and Human Services, Bethesda, MD (1987)

Kang AS, Barbas CF, Janda KD, Benkovic SJ, Lerner RA: Linkage of recognition and replication functions by assembling combinatorial antibody Fab libraries along phage surfaces. *Proc Natl Acad Sci USA* 88(10):4363-4366, 1991.

Kettleborough CA, Ansell KH, Allen RW, Rosell-Vives E, Gussow DH, Bendig MM: Isolation of tumor cell-specific single-chain Fv from immunized mice using phage-antibody libraries and the re-construction of whole antibodies from these antibody fragments. *Eur J Immunol* 24(4):952-958, 1994.

Kruger DH, Barcak GJ, Reuter M, Smith HO: EcoRII can be activated to cleave refractory DNA recognition sites. *Nucleic Acids Res* 16(9):3997-4008, (May 11) 1988.

Lalo D, Carles C, Sentenac A, Thuriaux P: Interactions between three common subunits of yeast RNA polymerases I and III. *Proc Natl Acad Sci USA* 90(12):5524-5528, 1993.

Laskowski M Sr: Purification and properties of venom phosphodiesterase. *Methods Enzymol* 65(1):276-84, 1980.

Lefkovits I and Pernis B, Editors. Immunological Methods, Vols. I and II. Academic Press, New York, NY. Also Vol. III published in Orlando and Vol. IV published in San Diego. ©1979-.

Lerner RA, Kang AS, Bain JD, Burton DR, Barbas CF 3d: Antibodies without immunization. *Science* 258(5086):1313-1314, 1992.

Leung, D.W., et al, *Technique*, 1:11-15, 1989.

Li B and Fields S: Identification of mutations in p53 that affect its binding to SV40 large T antigen by using the yeast two-hybrid system. *FASEB J* 7(10):957-963, 1993.

Lilley GG, Doelzal O, Hillyard CJ, Bernard C, Hudson PJ: Recombinant single-chain antibody peptide conjugates expressed in Escherichia coli for the rapid diagnosis of HIV. *J Immunol Methods* 171(2):211-226, 1994.

Lowman HB, Bass SH, Simpson N, Wells JA: Selecting high-affinity binding proteins by monovalent phage display. *Biochemistry* 30(45):10832-10838, 1991.

Luban J, Bossolt KL, Franke EK, Kalpana GV, Goff SP: Human immunodeficiency virus type 1 Gag protein binds to cyclophilins A and B. *Cell* 73(6):1067-1078, 1993.

Madura K, Dohmen RJ, Varshavsky A: N-recognin/Ubc2 interactions in the N-end rule pathway. *J Biol Chem* 268(16):12046-54, (Jun 5) 1993.

Marks JD, Griffiths Ad, Malmqvist M, Clackson TP, Bye JM, Winter G: By-passing immunization: building high affinity human antibodies by chain shuffling. *Biotechnology (N Y)* 10(7):779-783, 1992.

Marks JD, Hoogenboom HR, Bonnert TP, McCafferty J, Griffiths AD, Winter G: By-passing immunization. Human antibodies from V-gene libraries displayed on phage. *J Mol Biol* 222(3):581-597, 1991.

Marks JD, Hoogenboom HR, Griffiths AD, Winter G: Molecular evolution of proteins on filamentous phage. Mimicking the strategy of the immune system. *J Biol Chem* 267(23):16007-16010, 1992.

Maxam AM, Gilbert W: Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol* 65(1):499-560, 1980.

McCafferty J, Griffiths AD, Winter G, Chiswell DJ: Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* 348(6301):552-554, 1990.

Method of DNA sequencing.

Miller JH. A Short Course in Bacterial Genetics: A Laboratory Manual and Handbook for Escherichia coli and Related Bacteria (see inclusively p. 445). Cold Spring Harbor Laboratory Press, Plainview, NY, ©1992.

Milne GT and Weaver DT: Dominant negative alleles of RAD52 reveal a DNA repair/ recombination complex including Rad51 and Rad52. *Genes Dev* 7(9):1755-1765, 1993.

Mullinax RL, Gross EA, Amberg JR, Hay BN, Hogrefe HH, Kubtiz MM, Greener A, Alting-Mees M, Ardourel D, Short JM, et al: Identification of human antibody fragment clones specific for tetanus toxoid in a bacteriophage lambda immunoexpression library. *Proc natl Acad Sci USA* 87(20):8095-9099, 1990.

Nath K, Azzolina BA: in *Gene Amplification and Analysis* (ed. Chirikjian JG), vol. 1, p. 113, Elsevier North Holland, Inc., New York, New York, ©1981.

Needleman SB and Wunsch CD: A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48(3):443-453, 1970.

Nelson M, Christ C, Schildkraut I: Alteration of apparent restriction endonuclease recognition specificities by DNA methylases. *Nucleic Acids Res* 12(13):5165-73, 1984 (Jul 11).

Nicholls PJ, Johnson VG, Andrew SM, Hoogenboom HR, Raus JC, Youle RJ: Characterization of single-chain antibody (sFv)-toxin fusion proteins produced in vitro in rabbit reticulocyte lysate. *J Biol Chem* 268(7):5302-5308, 1993.

Oller AR, Vanden Broek W, Conrad M, Topal MD: Ability of DNA and spermidine to affect the activity of restriction endonucleases from several bacterial species. *Biochemistry* 30(9):2543-9, (Mar 5) 1991.

Owen MRL, Pen J: Transgenic Plants: A Production System for Industrial and Pharmaceutical Proteins. Chichester: John Wiley & Sons, 1996.

Owens RJ and Young RJ: The genetic engineering of monoclonal antibodies. *J Immunol Methods* 168(2):149-165, 1994.

Pearson WR and Lipman DJ: Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 85(8):2444-2448, 1988.

Pein CD, Reuter M, Meisel A, Cech D, Kruger DH: Activation of restriction endonuclease EcoRII does not depend on the cleavage of stimulator DNA. *Nucleic Acids Re s* 19(19):5139-42, (Oct 11) 1991.

Persson MA, Caothien RH, Burton DR: Generation of diverse high-affinity human monoclonal antibodies by repertoire cloning. *Proc Natl Acad Sci USA* 88(6):2432-2436, 1991.

Perun TJ, Propst CL, eds.: Computer-Aided Drug Design: Methods and Applications. New York: Marcel Dekker, Inc., 1989.

Qiang BQ, McClelland M, Poddar S, Spokauskas A, Nelson M: The apparent specificity of NotI (5'-GCGGCCGC-3') is enhanced by M.FnuDII or M.BepI methyltransferases (5'-mCGCG-3'): cutting bacterial chromosomes into a few large pieces. *Gene* 88(1):101-5, (Mar 30) 1990.

Queen C, Foster J, Stauber C, Stafford J: Cell-type specific regulation of a kappa immunoglobulin gene by promoter and enhance elements. *Immunol Rev* 89:49-68, 1986.

Raleigh EA, Wilson G: Escherichia coli K-12 restricts DNA containing 5-methylcytosine. *Proc Natl Acad Sci U S A* 83(23):9070-4, (Dec) 1986.

Reidhaar-Olson JF and Sauer RT: Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. *Science* 241(4861):53-57, 1988.

Riechmann L and Weill M: Phage display and selection of a site-directed randomized single-chain antibody Fv fragment for its affinity improvement. *Biochemistry*

32(34):8848-8855, 1993.

Roberts RJ, Macelis D: REBASE--restriction enzymes and methylases. *Nucleic Acids Res* 24(1):223-35, (Jan 1) 1996.

Ryan AJ, Royal CL, Hutchinson J, Shaw CH: Genomic sequence of a 12S seed storage protein from oilseed rape (Brassica napus c.v. jet neuf). *Nucl Acids Res* 17(9):3584, 1989.

Sambrook J, Fritsch EF, Maniatis T. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ©1982.

Sambrook J, Fritsch EF, Maniatis T. Molecular Cloning: A Laboratory Manual. Second Edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ©1989.

Scopes RK. Protein Purification: Principles and Practice. Springer-Verlag, New York, NY, © 1982.

Segel IH: Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems. New York: John Wiley & Sons, Inc., 1993.

Silver SC and Hunt SW 3d: Techniques for cloning cDNAs encoding interactive transcriptional regulatory proteins. *Mol Biol Rep* 17(3):155-165, 1993.

Smith TF, Waterman MS, Fitch WM: Comparative biosequence metrics. *J Mol Evol* S18(1):38-46, 1981.

Smith TF, Waterman MS. *Adv Appl Math* 2: 482-end of article, 1981.

Smith TF, Waterman MS: Identification of common molecular subsequences. *J Mol Biol* 147(1):195-7, (Mar 25) 1981.

Smith TF, Waterman MS: Overlapping genes and information theory. *J Theor Biol* 91(2):379-80, (Jul 21) 1981.

Staudinger J, Perry M, Elledge SJ, Olson EN: Interactions among vertebrate helix-loop-helix proteins in yeast using the two-hybrid system. *J Biol Chem* 268(7):4608-4611, 1993.

Stemmer WP, Morris SK, Wilson BS: Selection of an active single chain Fv antibody from a protein linker library prepared by enzymatic inverse PCR. *Biotechniques* 14(2):256-265, 1993.

Stemmer WP: DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc Natl Acad Sci USA* 91(22):10747-10751, 1994.

Sun D, Hurley LH: Effect of the (+)-CC-1065-(N3-adenine)DNA adduct on in vitro DNA synthesis mediated by Escherichia coli DNA polymerase. *Biochemistry* 31:10, 2822-9, (Mar 17) 1992,

Tague BW, Dickinson CD, Chrispeels MJ: A short domain of the plant vacuolar protein phytohemagglutinin targets invertase to the yeast vacuole. *Plant Cell* 2(6):533-46, (June) 1990.

Takahashi N, Kobayashi I: Evidence for the double-strand break repair model of bacteriophage lambda recombination. *Proc Natl Acad Sci U S A* 87(7):2790-4, (Apr) 1990.

Thiesen HJ and Bach C: Target Detection Assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. *Nucleic Acids Res* 18(11):3203-3209, 1990.

Thomas M, Davis RW: Studies on the cleavage of bacteriophage lambda DNA with EcoRI Restriction endonuclease. *J Mol Biol* 91(3):315-28, (Jan 25) 1975.

Tingey SV, Walker EL, Corruzzi GM: Glutamine synthetase genes of pea encode distinct polypeptides which are differentially expressed in leaves, roots and nodules.

EMBO J 6(1):1-9, 1987.

Topal MD, Thresher RJ, Conrad M, Griffith J: NaeI endonuclease binding to pBR322 DNA induces looping. *Biochemistry* 30(7):2006-10, (Feb. 19) 1991.

Tramontano A, Chothia C, Lesk AM: Framework residue 71 is a major determinant of the position and conformation of the second hypervariable region in the VH domains of immunoglobulins. *J Mol Biol* 215(1):175-182, 1990.

Tuerk C and Gold L: Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249(4968):505-510, 1990.

USPN 4,683,195; Filed Feb. 7, 1986, Issued Jul 28. 1987. Mullis KB, Erlich HA, Arnheim N, Horn GT, Saiki RK, Scharf SJ: Process for Amplifying, Detecting, and/or Cloning Nucleic Acid Sequences.

USPN 4,683,202; Filed Oct. 25, 1985, Issued Jul. 28, 1987. Mullis KB: Process for Amplifying Nucleic Acid Sequences.

USPN 4,704,362; Filed Nov. 5, 1979, Issued Nov. 3, 1987. Itakura K, Riggs AD: Recombinant Cloning Vehicle Microbial Polypeptide Expression.

USPN 4,713,337; Filed Jan. 3, 1985, Issued Dec. 15, 1987. Jasin M, Schimmel PR: Method for deletion of a gene from a bacteria.

USPN 4,732,856; Filed April 3, 1984, Issued March 22, 1988. Federoff NV: Transposable elements and process for using same.

USPN 4,963,487; Filed Sept. 14, 1987, Issued Jan. 16, 1990. Schimmel PR: Method for deletion of a gene from a bacteria.

USPN 5,354,656; Filed Oct. 2, 1989, Issued Oct. 11, 1994. Sorge, Joseph A. ; Huse, William D.:

USPN 5,385,835; Filed May 19, 1994, Issued Jan. 31, 1995. Helentjaris, Timothy ; Nienhuis, James: Identification and localization and introgression into plants of desired multigenic traits.

USPN 5,453,247; Filed Nov. 23, 1993, Issued Sept. 26, 1995. Beavis, Ronald C. ; Chait, Brian T.: Instrument and method for the sequencing of genome.

USPN 5,604,100; Filed July 19, 1995, Issued Feb. 18, 1997. Perlin, Mark W.: Method and system for sequencing genomes.

USPN 5,670,321; Filed May 10, 1995, Issued Sept. 23, 1997. Kimmel, Bruce E. ; Ellis, Michael ; Ruddy, David: Efficient method to conduct large-scale genome sequencing.

USPN 5,925,808; Filed Dec. 19, 1997, Issued July 20, 1999. Oliver, Melvin John ; Quisenberry, Jerry Edwin ; Trolinder, Norma Lee Glover ; Keim, Don Lee: Control Of Plant Gene Expression.

USPN 5,953,727; Filed March 6, 1997, Issued Sept. 14, 1999. Maslyn, Timothy J. ; Au-Young, Janice ; Hillman, Jennifer L. ; Hibbert, Harold ; Akerblom, Ingrid E. ; Cheng, Rachel J. ; Tang, Yuanhua T.:Project-based full-length biomolecular sequence database.

USPN 5,965,443; Filed Sept. 9, 1996, Issued Oct. 12, 1999. Reznikoff WS, Goryshin IY: System for in vitro transposition.

USPN 5,981,177; Filed Jan. 25, 1995, Issued Nov. 9, 1999. Demirjian DC, Casadaban MJ, Weber M, Gaines GL: Protein fusion method and constructs.

USPN 5,994,058; Filed March 20, 1995, Issued Nov. 30, 1999. Senapathy, Periannan:Method For Contiguous Genome Sequencing.

USPN 6,023,659; Filed March 6, 1997, Issued Feb. 8, 2000. Seilhamer, Jeffrey J. ; Akerblom, Ingrid E. ; Altus, Christina M. ; Klingler, Tod M. ; Russo, Frank ; Au-Young, Janice ; Hillman, Jennifer L. ; Maslyn, Timothy J.: Database System Employing Protein Function Hierarchies For Viewing Biomolecular Sequence Data.

van de Poll ML, Lafleur MV, van Gog F, Vrieling H, Meerman JH: N-acetylated and deacetylated 4'-fluoro-4-aminobiphenyl and 4-aminobiphenyl adducts differ in their ability to inhibit DNA replication of single-stranded M13 in vitro and of single-stranded phi X174 in Escherichia coli. *Carcinogenesis* 13(5):751-8, (May) 1992.

Vojtek AB, Hollenberg SM, Cooper JA: Mammalian Ras interacts directly with the serine/threonine kinase Raf. *Cell* 74(1):205-214, 1993.

Wenzler H, Mignery G, Fisher L, Park W: Sucrose-regulated expression of a chimeric potato tuber gene in leaves of transgenic tobacco plants. *Plant Mol Biol* 13(4):347-54, 1989.

White JS, White DC: Source Book of Enzymes. Boca Raton: CRC Press, 1997.

Williams and Barclay, in Immunoglobulin Genes, The Immunoglobulin Gene Superfamily

Winnacker EL. From Genes to Clones: Introduction to Gene Technology. VCH Publishers, New York, NY, ©1987.

Winter G and Milstein C: Man-made antibodies. *Nature* 349(6307):293-299, 1991.

WO 00/04190; Filed July 15, 1999, Published Jan. 27, 2000. Del Cardayre S, Tobin M, Stemmer WP, Ness JE, Minshull J, Patten PA, Subramanian V, Castle LA, Krebber CM, Bass S, Zhang Y, Cox T, Huisman G, Yuan L, Affholter JA: Evolution of whole cells and organisms by recursive sequence recombination.

WO 00/09755; Filed Aug. 12, 1999, Published Feb. 24, 2000. Zarling D, Reddy G, Pati S: Domain specific gene evolution.

WO 88/08453; Filed Apr. 14, 1988, Published Nov. 3, 1988. Alakhov JB, Baranov, VI, Ovodov SJ, Ryabova LA, Spirin AS: Method of Obtaining Polypeptides in Cell-Free Translation System.

WO 90/05785; Filed Nov. 15, 1989, Published May 31, 1990. Schultz P: Method for Site-Specifically Incorporating Unnatural Amino Acids into Proteins.

WO 90/07003; Filed Jan. 27, 1989, Published June 28, 1990. Baranov VI, Morozov IJ, Spirin AS: Method for Preparative Expression of Genes in a Cell-free System of Conjugated Transcription/translation.

WO 91/02076; Filed June 14, 1990, Published Feb. 21, 1991. Baranov VI, Ryabova LA, Yarchuk OB, Spirin AS: Method for Obtaining Polypeptides in a Cell-free System.

WO 91/05058; Filed Oct. 5, 1989, Published Apr. 18, 1991. Kawasaki G: Cell-free Synthesis and Isolation of Novel Genes and Polypeptides.

WO 91/17271; Filed May 1, 1990, Published Nov. 14, 1991. Dower WJ, Cwirla SE: Recombinant Library Screening Methods.

WO 91/18980; Filed May 13, 1991, Published Dec. 12, 1991. Devlin JJ: Compositions and Methods for Indentifying Biologically Active Molecules.

WO 91/19818; Filed June 20, 1990, Published Dec. 26, 1991. Dower WJ, Cwirla SE, Barrett RW: Peptide Library and Screening Systems.

WO 92/02536; Filed Aug. 1, 1991, Published Feb. 20, 1992. Gold L, Tuerk C: Systematic Polypeptide Evolution by Reverse Translation.

WO 92/03918; Filed Aug. 28, 1991, Published Mar. 19, 1992. Lonberg N, Kay RM: Transgenic Non-human Animals Capable of Producing Heterologous Antibodies.

WO 92/05258; Filed Sept. 17, 1991, Published Apr. 2, 1992. Fincher GB: Gene Encoding Barley Enzyme.

WO 92/14843; Filed Feb. 21, 1992, Published Sept. 3, 1992. Toole JJ, Griffin LC, Bock LC, Latham JA, Muenchau DD, Krawczyk S: Aptamers Specific for Biomolecules and Method of Making.

WO 93/08278; Filed Oct. 15, 1992, Published Apr. 29, 1993. Schatz PJ, Cull MG, Miller JF, Stemmer WP: Peptide Library and Screening Method.

WO 93/12227; Filed Dec. 17, 1992, Published June 24, 1993. Lonberg N, Kay RM: Transgenic Non-human Animals Capable of Producing Heterologous Antibodies.

WO 94/25585; Filed Apr. 25, 1994, Published Nov. 10, 1994. Lonberg N, Kay RM: Transgenic Non-human Animals Capable of Producing Heterologous Antibodies.

WO 95/00530; Filed June 6, 1994, Published Jan. 1, 1995. Fodor, Stephen, P., A. ; Lipshutz, Robert, J. ; Huang, Xiaohua ; Jevons, Luis, Carlos: Hybridization and Sequencing of Nucleic Acids.

WO 96/21031; Filed June 7, 1995, Published July 11, 1996. Tricoli, David, M. ; Carney, Kim, J. ; Russell, Paul, F. ; Quemada, Hector, D. ; Mcmaster, J., Russell ; Reynolds, John, F. ; Deng, Rosaline, Z.: Transgenic Plants Expressing DNA Constructs Containing A Plurality Of Genes To Impart Virus Resistance.

WO 96/27025; Filed Feb. 21, 1996, Published Sept. 6, 1996. Rabani, Ely, Michael:Device, Compounds, Algorithms, And Methods Of Molecular Characterization And Manipulation With Molecular Parallelism.

WO 97/17429; Filed Nov. 8, 1996, Published May 15, 1997. Oglevee-O'donovan, Wendy ; Arteca, Richard, N. ; Arteca, Jeannette ; Stoots, Eleanor: Method For The Commercial Production Of Transgenic Plants.

WO 97/35966; Filed March 20, 1997, Published Oct. 2, 1997. Minshull J, Stemmer WP: Methods and compositions for cellular and metabolic engineering.

WO 97/37041; Filed March 18, 1997, Published Oct. 9, 1997. Köster, Hubert: DNA Sequencing By Mass Spectrometry.

WO 97/42348; Filed May 5, 1997, Published Nov. 13, 1997. Köster, Hubert ; Van Den Boom, Dirk ; Ruppert, Andreas: Process For Direct Sequencing During Template Amplification.

WO 98/26407; Filed Dec. 11, 1997, Published June 18, 1998. Sabatini, Cathryn, E. ; Heath, Joe, Don ; Covitz, Peter, A. ; Klinger, Tod, M. ; Russo, Frank, D. ; Berry, Stephanie, F.: Database And System For Storing, Comparing And Displaying Genomic Information.

WO 98/26408; Filed Dec. 11, 1997, Published June 18, 1998. Sabatini, Cathryn, E. ; Heath, Joe, Don ; Covitz, Peter, A. ; Klingler, Tod, M. ; Russo, Frank, D. ; Berry, Stephanie, F.:Database And System For Determining, Storing And Displaying Gene Locus Information.

WO 98/31833; Filed Dec. 12, 1997, Published July 23, 1998. Ju, Jingyue: Nucleic Acid Sequencing With Solid Phase Capturable Terminators.

WO 98/31834; Filed Dec. 12, 1997, Published July 23, 1998. Ju, Jingyue: Sets Of Labeled Energy Transfer Fluorescent Primers And Their Use In Multi Component Analysis.

WO 98/31837; Filed Jan. 16, 1998, Published July 23, 1998. Delcardayre SB, Tobin MB, Stemmer WP, Ness JE, Minshull J, Patten P: Evolution of whole cells and organisms by recursive sequence recombination.

WO 98/36085; Filed Feb. 13, 1998, Published Aug. 20, 1998. Sutliff, Thomas, D. ; Rodriguez, Raymond, L.: Production Of Mature Proteins In Plants.

WO 98/37223; Filed Feb. 18, 1998, Published Aug. 27, 1998. Pang, Sheng-Zhi ; Gonsalves, Dennis ; Jan, Fuh-Jyh: DNA Construct To Confer Multiple Traits On Plants.

WO 99/35494; Filed Jan. 8, 1999, Published July 15, 1999. Tally FP, Tao J, Wendler PA, Connelly G, Gallant PL: Method for identifying validated target and assay combinations.

WO 99/37755; Filed Dec. 11, 1998, Published July 29, 1999. Pati S, Zarling David, Lehman CW, Zeng H: The use of consensus sequences for targeted homologous gene isolation and recombination in gene families.

WO 99/49403; Filed March 25, 1999, Published Sept. 30, 1999. Lincoln, Stephen, E. ; Hodgson, David, M. ; Spiro, Peter, A. ; Russo, Frank, D. ; Akerblom, Ingrid, E. ; Hillman, Jennifer, L. ; Jones, Anissa, Lee ; Bratcher, Shawn, Robert ; Cohen, Howard, Jerome ; Dufour, Gerard ; Wood, Michael, Peter ; Koleszar, Alexander, George ; Banville, Steven, C.: System And Methods For Analyzing Biomolecular Sequences.

WO95/11995; Filed Oct. 26, 1994, Published May 4, 1995. Chee M, Cronin MT, Fodor SP, Gingeras TR, Huang XC, Hubbell EA, Lipshutz RJ, Lobban PE, Miyada CG, Morris MS, Shah N, Sheldon EL: Arrays Of Nucleic Acid Probes On Biological Chips.

Wong CH, Whitesides GM: Enzymes in Synthetic Organic Chemistry. Vol. 12. New York: Elsevier Science Publications, 1995.

Yang X, Hubbard EJ, Carlson M: A protein kinase substrate identified by the two-hybrid system. *Science* 257(5070):680-2, (Jul 31) 1992.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R.: Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 17(10):994-9 (Oct) 1999.

Hopkins MJ, Sharp R, Macfarlane GT.: Age and disease related changes in intestinal bacterial populations assessed by cell culture, 16S rRNA abundance, and community cellular fatty acid profiles. *Gut* 48(2):198-205 (Feb) 2001.

Ritchie NJ, Schutter ME, Dick RP, Myrold DD.: Use of length heterogeneity PCR and fatty acid methyl ester profiles to characterize microbial communities in soil*Appl Environ Microbiol* 66(4):1668-75 (Apr) 2000.

Khan AA, Wang RF, Cao WW, Franklin W, Cerniglia CE.: Reclassification of a polycyclic aromatic hydrocarbon-metabolizing bacterium, Beijerinckia sp. strain B1, as Sphingomonas yanoikuyae by fatty acid analysis, protein pattern analysis, DNA-DNA hybridization, and 16S ribosomal DNA sequencing. *Int J Syst Bacteriol* 46(2):466-9 (Apr) 1996.

Peltroche-Llacsahuanga H, Schmidt S, Lutticken R, Haase G.: Discriminative power of fatty acid methyl ester (FAME) analysis using the microbial identification system (MIS) for Candida (Torulopsis) glabrata and Saccharomyces cerevisiae. *Diagn Microbiol Infect Dis* 38(4):213-21 (Dec) 2000.

SA Gerber et al.: Analysis of rates of multiple enzymes in cell lysates by electrospray ionization mass spectrometry. *J. Am. Chem. Soc.* 121:1102-3 1999.

www.genomeweb.com
David Goodlett discusses the latest in genomics – ICAT reagents
Written by: Marian Moser Jones
Dec 20, 2000

WO0011208; Filed Aug 25, 1999, Published March 2, 2000. Aebersold RH, Gelb MH, Gygi, SP, Scott CR, Turecek F, Gerber SA, Rist B: Rapid quantitative analysis of proteins or protein function in complex mixtures.

WO9905221; Filed July 27 1998, Published Feb. 4,1999. Cummins WJ, West RM, Smith JA: Cyanine Dyes.

US4876350; Filed Dec 16, 1987, Issued Oct 24, 1989. McGarrity J, Tenud L: Process for the production of (+) biotin.

US5776723; Filed Feb 8, 1996, Issued July 7, 1998. Herold CD, O'Hagan M: Rapid detection of mycobacterium tuberculosis.

US6136173; Filed June 24, 1996, Issued Oct. 24, 2000. Anderson NL, Anderson NG, Goodman J: Automated system for two-dimensional electrophoresis.

US6127134; Filed April 20, 1995, Issued Oct. 3, 2000. Minden J, Waggoner A:

Difference gel electrophoresis using matched multiple dyes.
US6064754; Filed Dec 1, 1997, Issued May 16, 2000. Parekh RB, Amess R, Bruce JA, Prime SB, Platt AE, Stoney RM: Computer-assisted methods and apparatus for identification and characterization of biomolecules in a biological sample.
US6013165; Filed May 22, 1998, Issued Jan 11,2000. Wiktorowicz JE, Raysberg Y: Electrophoresis apparatus and method.
Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K Editors.Current Protocols In Molecular Biology, Vol 2. John Wiley & Sons, Inc, © 2001, 10.21.4-10.21.6, 10.22.5-10.22.10, 10.22.14, 10.22.15-10.22.20.
Sambrook J, Russell DW Editors. Molecular Cloning A Laboratory Manual 3rd ed. Cold Spring Harbor Laboratory Press, New York, © 2001, 18.3, 18.62, 18.66.

1.4.8. Additional methods for differential analysis

1.4.8.1. Protein expression profiling using selective differential labeling

The use of mass spectrometry to identify proteins whose sequences are present in either DNA or protein databases is well established and integral to the field of Proteomics. Protein and peptide mass can be determined at high accuracy by several mass spectrometric techniques. Peptide can be further fragmented in a tandem or ion trap mass spectrometer yielding sequence information of the peptide. Both types of mass information can be used to identify protein in a sequence database. One goal of Proteomics is to define the expressed proteins associated with a given cellular state and another is to quantify changes in protein expression between cellular states. One of the new methodologies that have a great impact on proteome research is known as isotope-coded affinity tag (ICAT) peptide labeling (17). The method is based on a newly synthesized class of chemical reagents (ICATs) used in combination with tandem mass spectrometry. The ICAT reagent contains a biotin affinity tag and a thiol specific reactive group, which are joined by a spacer domain which is available in two forms: regular and isotopically heavy, which includes eight deuterium atoms. First, a reduced protein mixture representing one cell state is derivatized with the isotopically light version of the ICAT reagent, while the corresponding reduced protein mixture representing a second cell state is derivatized with the isotopically heavy version of the ICAT reagent. Second, the labeled samples are combined and proteolytically digested to produce peptide fragments. Third, the tagged cysteine containing peptide fragments are isolated by avidin affinity chromatography. Finally, the isolated tagged peptides are separated and analyzed by microcapillary tandem mass spectrometry.

There are, however, limitations associated with their approach: (i) differential labeling reagents relied on stable isotopes which is expensive and not very flexible to multiplex differential labeling; (ii) The moieties attached to the original peptides are approximately 500 Dalton heavy, which is heavier than some peptides and is likely to affect peptide ionization and fragmentation process; (iii) Some bonds in the labeling reagent are week compared to the amide bond, which might complicate the MS/MS spectrum, (iv) Protein expression profiling is limited to duplex comparison; (v) The affinity interaction between biotin and avidin is too strong to release the immobilized peptide efficiently.

In one embodiment, this present invention provides a method for simultaneous identification and quantification of expression levels of individual proteins carrying certain functional groups in their side chains. The proteins may be analyzed in complex mixtures. The method is based on comparison of two or more samples of proteins, one of which can be considered as the standard sample and all others can be considered as samples under investigation.

The samples of proteins are subjected to a sequence of manipulations including (i) proteolytic digestion into mixtures of peptides, (ii) treatment of the mixtures of peptides with chemical probes, (iii) washing away and discarding the unbound peptides from the mixtures, (iv) cleaving the chemical probes and the consequential release of the peptides still carrying parts of the chemical probes into solution. This sequence of manipulations may also include one or more auxiliary chemical and/or enzymatic modifications of functional groups in side chains and/or in the free termini of the proteins and/or peptides in order to achieve selective and the most favorable modification for the next steps in the protocol. The auxiliary modifications may be performed between any steps of the main sequence.

The core structure of the chemical probe consists of (i) a solid support, (ii) a spacer, (iii) a cleavable moiety, (iv) a differential mass labeling unit, and (v) a reactive group. The chemical probes perform three functions: (i) they attach peptides carrying specific functional groups in their side chains and/or termini to a solid support by forming covalent chemical bonds to the reactive group of the probe, (ii) they provide means for selective cleavage of the attached peptide from the solid support such that a part of the probe still remains attached to the peptide, and (iii) they serve as differential labeling reagents.

Differential labeling results from attaching of chemical moieties of different mass but of similar properties to a protein or a peptide such that peptides with the same sequence but with different labels are eluted together in the separation procedure and their ionization and detection properties regarding mass spectrometrical analysis are very similar. The differential mass labeling unit remains covalently bound to the peptide after it is cleaved from the solid support part of the probe. Signals corresponding to peptides with the same sequence but marked with differential mass labels are assigned to different original protein samples.

The auxiliary chemical and/or enzymatic modification can be used to introduce additional differential mass labels into the peptides.

The reactive group on the chemical probe may be activated or modified by a bridging reagent prior to a reaction with mixtures of peptides. Such activation or modification provides for a greater flexibility in design of the chemical probe since the same core structure of a chemical probe may be tuned to increase reactivity and/or selectivity towards different functional groups in side chains and/or in termini of the peptides.

After being cleaved from the solid support part of the chemical probe, the differentially labeled peptide mixtures are combined, subjected to multidimensional chromatographic separation, and analyzed by mass spectrometry methods. Mass spectrometry data is processed by special software, which allows for determination and tracing the composition and sequence of peptides in the mixture to identification of the original proteins and their quantification.

This approach can be used for duplex or potentially multiplex protein expression profiling. The complexity of the sample is simplified by targeting peptides containing particular amino acids, which selected by a reaction with chemical probes.

Novelties of this invention include: (i) design of solid phase-based differential mass labeling reagents for selective peptide modification; (ii) design of various kinds of differential mass unit; (iii) combination of differential mass probes with various bridge reagent to target certain amino acid specifically; (iv) multiplex analysis; (v) combination of proteolytic digestion and chemical and/or enzymatic modifications in side chains and/or in termini of proteins and peptides in order to achieve selective and the most favorable modifications for the next steps in the protocol; (vi) combination of differential chemical labeling with MudPIT, and possible all other protein/peptide separation or purification technologies if necessary.

One embodiment of this invention provides reagents and procedures for quantification of protein expression using combination of selective differential peptides labeling, and LC MS/MS or LC-LC MS/MS. This invention overcomes the limitations inherent in traditional techniques. The basic approach described can be employed for quantitative analysis of protein expression in complex samples (such as cells, tissues, and fraction etc.), the detection and quantitation of specific proteins in complex samples, and quantitative measurement of specific enzymatic activities in complexed samples.

1.4.8.2. Technical description

1. Probe design:

The solid support part of the chemical probe may consist of any of the following materials or any combination of them: gel, glass beads, magnetic beads, polymers, silicon wafer, membrane, or resin.

The spacer between the solid phase part and the cleavable unit of the chemical probe may be included for convenience and improved yields in synthetic preparation of the chemical probe. The spacer may consist of a chain of 2 to 8 atoms, which can be C, O, N, B, Si, S, P, Se …, covalently bound to each other. In order to satisfy the valence requirements, the atoms may carry hydrogen atoms, halogens, or one of the following groups containing up to 25 atoms: alkyl, hydroxy, alkoxy, amino, alkylamino... The spacer may contain cyclic moieties with or without heteroatoms and with or without substituents.

The cleavable moiety provides means for selective detachment of the solid phase part of the chemical probe from the differential mass label attached to peptide. It is designed such that it can be cleaved by treating the probe with a chemical reagent or any kind of electromagnetic irradiation, photochemically, enzymatically, or thermally.

Differential mass labeling units differ in molecular mass, but do not differ in retention properties regarding the separation method used and in ionization and detection properties regarding the mass spectrometry methods used. These moieties differ either in their isotope composition (isotopic labels) or they differ structurally by a rather small fragment, which change does not alter the properties stated above (homologous labels).

The isotopic labels can be presented by general formulae:

Z^A and Z^B

Z^A and Z^B = R-Z^1-A^1-Z^2-A^2-Z^3-A^3-Z^4-A^4-

Z^1, Z^2, Z^3, and Z^4 independently of one another can be selected from O, OC (O), OC (S), OC (O) O, OC (O) NR, OC (S) NR, OSiRR^1, S, SC (O), SC (S), SS, S (O), S (O_2), NR, NRR^{1+}, C (O), C (O) O, C (S), C (S) O, C (O) S, C (O) NR, C (S) NR, SiRR^1, (Si (RR^1) O) n, SnRR^1, Sn (RR^1) O, BR (OR^1), BRR^1, B (OR)(OR^1), OBR (OR^1), OBRR^1, OB (OR)(OR^1) or Z^1 - Z^4 may be absent;

A^1, A^2, A^3, and A^4 independently of one another can be selected from (CRR^1)n, in which some single C-C bonds may be replaced with double or triple bonds, in which case some groups R and R^1 will be absent, *o*-arylene, *m*-arylene, *p*-arylene with up to 6 substituents, carbocyclic, bicyclic, or tricyclic fragments with up to 8 atoms in the cycle with or without heteroatoms (O, N, S) and with or without substituents, or A^1 - A^4 may be absent;

R, R^1 independently from other R and R^{1}in Z^1 - Z^4 and independently from other R and R^1 in A^1 - A^4 is hydrogen, halogen, an alkyl, alkenyl, alkynyl, or aryl group;

n in Z^1 - Z^4 is independent of n in A^1 - A^4 and is a whole number that can have value from 0 to 21.

Z^A has the same structure as Z^B, but they have different isotope composition. For instance, if Z^A contains *x* number of protons, Z^B may contain *y* number of deuterons in the place of protons, and, correspondingly, *x* - *y* number of protons remaining; and/or if Z^A contains *x* number of borons-10, Z^B may contain *y* number of borons-11 in the place of borons-10, and, correspondingly, *x* - *y* number of borons-10 remaining; and/or if Z^A contains *x* number of carbons-12, Z^B may contain *y* number of carbons-13 in the place of carbons-12, and, correspondingly, *x* - *y* number of carbons-12 remaining; and/or if Z^A contains *x* number of nitrogens-14, Z^B may contain *y* number of nitrogens-15 in the place of nitrogens-14, and, correspondingly, *x* - *y* number of nitrogens-14 remaining; and/or if Z^A contains *x* number of sulfurs-32, Z^B may contain *y* number of sulfurs-34 in the place of sulfurs-32, and, correspondingly, *x* - *y* number of sulfurs-32 remaining; and so on for all elements which may be present and have different stable isotopes.

x and *y* are whole numbers between 1 and 21 such that *x* is greater than *y*.

An example of an isotopical label pairs/series: $(CD_2)_n$ / $(CH_2)_n$, where n = 0, 1, 2, ..., 21; (delta mass = 2n)

The homologous reagents can be presented by general formulae:

Z^A and Z^B where Z^A and Z^B = $R\text{-}Z^1\text{-}A^1\text{-}Z^2\text{-}A^2\text{-}Z^3\text{-}A^3\text{-}Z^4\text{-}A^4\text{-}$

Z^1, Z^2, Z^3, and Z^4 independently of one another can be selected from O, OC(O), OC(S), OC(O)O, OC(O)NR, OC(S)NR, OSiRR^1, S, SC(O), SC(S), SS, S(O), S(O_2), NR, NRR^{1+}, C(O), C(O)O, C(S), C(S)O, C(O)S, C(O)NR, C(S)NR, SiRR^1, (Si(RR^1)O)n, SnRR^1, Sn(RR^1)O, BR(OR^1), BRR^1, B(OR)(OR^1) , OBR(OR^1), OBRR^1, OB(OR)(OR^1) or Z^1 - Z^4 may be absent;

A^1, A^2, A^3, and A^4 independently of one another can be selected from (CRR^1)n, in which some single C-C bonds may be replaced with double or triple bonds, in which case some groups R and R^1 will be absent, *o*-arylene, *m*-arylene, *p*-arylene with up to 6 substituents, carbocyclic, bicyclic, or tricyclic fragments with up to 8 atoms in the cycle with or without heteroatoms (O, N, S) and with or without substituents, or A^1 - A^4 may be absent;

R, R^1 independently from other R and R^1in Z^1 - Z^4 and independently from other R and R^1 in A^1 - A^4 is hydrogen, halogen, an alkyl, alkenyl, alkynyl, or aryl group;

n in Z^1 - Z^4 is independent of n in A^1 - A^4 and is a whole number that can have value from 0 to 21.

Z^A has a similar structure to that of Z^B, but Z^A has *x* extra $-CH_2-$ fragment(s) in one or more A^1 - A^4 fragments, and/or Z^A has *x* extra $-CF_2-$ fragment(s) in one or more A^1 - A^4 fragments; and/or if Z^A contains *x* number of protons, Z^B may contain *y* number of halogens in the place of protons, and, correspondingly, *x* - *y* number of protons remaining in one or more A^1 - A^4 fragments; and/or Z^A has *x* extra –O- fragment(s) in one or more A^1 - A^4 fragments; and/or Z^A has *x* extra –S- fragment(s) in one or more A^1 - A^4 fragments; and/or if Z^A contains *x* number of –O- fragment(s), Z^B may contain *y* number of –S- fragment(s) in the place of –O- fragment(s), and, correspondingly, *x* - *y* number of –O- fragment(s) remaining in one or more A^1 - A^4 fragments; and so on.

x and *y* are whole numbers between 1 and 21 such that *x* is greater than *y*.

An examples of homologous label pairs/series: $(CH_2)_n/(CH_2)_{n+m}$, where n = 0, 1, 2, …, 21; m = 1, 2, …, 21 (delta mass = 14m)

2. Bridging and activating reagents: We may either utilize some commercial available cross linkers or synthesized our own:

a. Reactive site 1: probe specific

b. Reactive site 2: amino acid specific

3. Methods for peptide/protein separation and detection:

On line 2 dimensional capillary LC ESI MS/MS (MuDPIT) as described in the global differential profiling disclosure, or 1 D LC ESI MS/MS, MALDI MS.

4. Sequence analysis and quantification:

Peptides are quantified by measuring in the MS mode the relative signal intensities for pairs or series of peptide ions of identical sequence that are tagged differentially, which therefore differ in mass by the mass differential encoded within the differential labeling reagents. Peptide sequence information is automatically generated by selecting peptide ions of a particular mass-to-charge (m/z) ratio for collision-induced dissociation (CID) in the mass spectrometer operating in the tandem MS mode. (Link *et al, Electrophoresis* 18:1314-34 (1997); Gygi *et al. Nature Biotechnol* 17:994-9) (1999); Gygi *et al., cell Biol* 19:1720-30 (1999)).

The resulting tandem mass spectra can be correlated to sequence databases to identify the protein from which the sequenced peptide originated. Currently commercial available softwares are Turbo SEQUEST by Thermofinigan, MassScot by Matrix Science, and Sonar MS/MS by Proteometrics. Special software development will be necessary for automated relative quantification.

One suggested approach of practicing the invention:

1. Protein sample preparation, which may include protein denaturation, reduction, and proteolytic digestion
2. Treatment of the probe with a desired activating or bridging reagent
3. Treatment of the activated probe with a mixture of peptides
4. Wash off unbound peptides, which don't have the targeted amino acid
5. Combining modified differential labeled peptide mixture
6. Release peptides by cleaving the probe (steps 5 and 6 can be switched)
7. Removing solvent or desalting if necessary
8. Redisovling peptide in LC loading buffer
9. LC ESI MS and MS/MS analysis MALDI MS and MS/MS analysis
10. Database searching and data analysis

1.5. Metabolomics and lipidomics

Additional holistic monitoring approaches, metabolomics and lipidomics, include profiling metabolite pools, carbohydrates, lipids, glycoproteins, and glycolipids Various chromatographic methods and other qualitative and/or quantitative methods could be utilized to characterize lipid profiles. In the area of metabolomics, methods that compare concentrations of metabolites/small molecules, using a variety of chemical analysis tools, e.g. mass spec, NMR, other spectroscopic techniques, biosensors could be utilized. For some specific method examples, see the following references: J. C. Lindon et al., Prog. NMR Spear., 29, 1 (1996)l- J. C. Lindon et al., Drug. Met. Rev., 29, 705 (1997); B. Vogler et al., J Nat. Prod., 61, 175 (1998); and JA. Wolfender et al., Curr. Org. Chem. 2, 575 (1998); J. K. Nicholson et al., Xenobiotica, 29, 1181(1999).

1.6. Screening tools

1.6.1. FACS

Fluorescence activated cell sorting (FACS) methods are also a powerful tool for selection/screening. In some instances a fluorescent molecule is made within a cell (e.g., green fluorescent protein). The cells producing the protein can simply be sorted by FACS. Gel microdrop technology allows screening of cells encapsulated in agarose microdrops (Weaver et al. Methods 2:234-247 (1991)). In this technique products secreted by the cell (such as antibodies or antigens) are immobilized with the cell that generated them. Sorting and collection of the drops containing the desired product thus also collects the cells that made the product, and provides a ready source for the cloning of the genes encoding the desired functions. Desired products can be detected by incubating the encapsulated cells with fluorescent antibodies (Powell et al. Bio/Technology 8:333-337 (1990)). FACS sorting can also be used by this technique to assay resistance to toxic compounds and antibiotics by selecting droplets that contain multiple cells (i.e., the product of continued division in the presence of a cytotoxic compound; Goguen et al. Nature 363:189-190 (1995)). This method can select for any enzyme that can change the fluorescence of a substrate that can be immobilized in the agarose droplet.

1.6.2. Reporter molecule

In some embodiments of the invention, screening can be accomplished by assaying reactivity with a reporter molecule reactive with a desired feature of, for example, a gene product. Thus, specific functionalities such as antigenic domains can be screened with antibodies specific for those determinants.

1.6.3. Cell-cell indicator

In other embodiments of the invention, screening is preferably done with a cell-cell indicator assay. In this assay format, separate library cells (Cell A, the cell being assayed) and reporter cells (Cell B, the assay cell) are used.

Only one component of the system, the library cells, is allowed to evolve. The screening is generally carried out in a two-dimensional immobilized format, such as on plates. The products of the metabolic pathways encoded by these genes (in this case, usually secondary metabolites such as antibiotics, polyketides, carotenoids, etc.) diffuse out of the library cell to the reporter cell. The product of the library cell may affect the reporter cell in one of a number of ways.

The assay system (indicator cell) can have a simple readout (e.g., green fluorescent protein, luciferase, beta- galactosidase) which is induced by the library cell product but which does not affect the library cell. In these examples the desired product can be detected by colorimetric changes in the reporter cells adjacent to the library cell.

1.6.4. Feedback mechanism

In other embodiments, indicator cells can in turn produce something that modifies the growth rate of the library cells via a feedback mechanism. Growth rate feedback can detect and accumulate very small differences. For example, if the library and reporter cells are competing for nutrients, library cells producing compounds to inhibit the growth of the reporter cells will have more available nutrients, and thus will have more opportunity for growth. This is a useful screen for antibiotics or a library of polyketide synthesis gene clusters where each of the library cells is expressing and exporting a different polyketide gene product.

1.6.5. Screening Secreted molecules

Another variation of this theme is that the reporter cell for an antibiotic selection can itself secrete a toxin or antibiotic that inhibits growth of the library cell. Production by the library cell of an antibiotic that is able to suppress growth of the reporter cell will thus allow uninhibited growth of the library cell.

Conversely, if the library is being screened for production of a compound that stimulates the growth of the reporter cell (for example, in improving chemical syntheses, the library cell may supply nutrients such as amino acids to an auxotrophic reporter, or growth factors to a growth-factor- dependent reporter. The reporter cell in turn should produce a compound that stimulates the growth of the library cell. Interleukins, growth factors, and nutrients are possibilities. Further possibilities include competition based on ability to kill surrounding cells, positive feedback loops in which the desired product made by the evolved cell stimulates the indicator cell to produce a positive growth factor for cell A, thus indirectly selecting for increased product formation.

In some embodiments of the invention it can be advantageous to use a different organism (or genetic background) for screening than the one that will be used in the final product. For example, markers can be added to DNA constructs used for recursive sequence recombination to make the microorganism dependent on the constructs during the improvement process, even though those markers may be undesirable in the final recombinant microorganism.

Likewise, in some embodiments it is advantageous to use a different substrate for screening an evolved enzyme than the one that will be used in the final product. For example, Evnin et al. (Proc. Natl. Acad. Sci. U.S.A. 87:6659-6663 (1990)) selected trypsin variants with altered substrate specificity by requiring that variant trypsin generate an essential amino acid for an arginine auxotroph by cleaving arginine beta-naphthylamide. This is thus a selection for arginine-specific trypsin, with the growth rate of the host being proportional to that of the enzyme activity.

The pool of cells surviving screening and/or selection is enriched for recombinant genes conferring the desired phenotype (e.g. altered substrate specificity, altered biosynthetic ability, etc.). Further enrichment can be obtained, if desired, by performing a second round of screening and/or selection without generating additional diversity.

The recombinant gene or pool of such genes surviving one round of screening/selection forms one or more of the substrates for a second round of recombination. Again, recombination can be performed in vivo or in vitro by any of the recursive sequence recombination formats described above.

If recursive sequence recombination is performed in vitro, the recombinant gene or genes to form the substrate for recombination should be extracted from the cells in which screening/selection was performed. Optionally, a subsequence of such gene or genes can be excised for more targeted subsequent recombination. If the recombinant gene(s) are contained within episomes, their isolation presents no difficulties. If the recombinant genes are chromosomally integrated, they can be isolated by amplification primed from known sequences flanking the regions in which recombination has occurred. Alternatively, whole genomic DNA can be isolated, optionally amplified, and used as the substrate for recombination. Small samples of genomic DNA can be amplified by whole genome amplification with degenerate primers (Barrett et al. Nucleic Acids Research 23:3488-3492 (1995)). These primers result in a large amount of random 3' ends, which can undergo homologous recombination when reintroduced into cells.

If the second round of recombination is to be performed in vivo, as is often the case, it can be performed in the cell surviving screening/selection, or the recombinant genes can be transferred to another cell type (e.g., a cell is type having a high frequency of mutation and/or recombination). In this situation, recombination can be effected by introducing additional DNA segment(s) into cells bearing the recombinant genes. In other methods, the cells can be induced to exchange genetic information with each other by, for example, electroporation. In some methods, the second round of recombination is performed by dividing a pool of cells surviving screening/selection in the first round into two subpopulations. DNA from one subpopulation is isolated and transfected into the other population, where the recombinant gene(s) from the two subpopulations recombine to form a further library of recombinant genes. In these methods, it is not necessary to isolate particular genes from the first subpopulation or to take steps to avoid random shearing of DNA during extraction. Rather, the whole genome of DNA sheared or otherwise cleaved into manageable sized fragments is transfected into the second subpopulation. This approach is particularly useful when several genes are being evolved simultaneously and/or the location and identity of such genes within chromosome are not known.

The second round of recombination is sometimes performed exclusively among the recombinant molecules surviving selection. However, in other embodiments, additional substrates can be introduced. The additional substrates can be of the same form as the substrates used in the first round of recombination, i.e., additional natural or induced mutants of the gene or cluster of genes, forming the substrates for the first round. Alternatively, the additional substrate(s) in the second round of recombination can be exactly the same as the substrate(s) in the first round of replication.

After the second round of recombination, recombinant genes conferring the desired phenotype are again selected. The selection process proceeds essentially as before. If a suicide vector bearing a selective marker was used in the first round of selection, the same vector can be used again. Again, a cell or pool of cells surviving selection is selected. If a pool of cells, the cells can be subject to further enrichment.

1.4. Screening for various potential applications

1.4.1 Novel drugs: identifying targets

The invention relates to procedures that can be applied to identifying compounds that bind to and modulate the function of target components of a cell whose function is known or unknown, and cell components that are not amenable to other screening methods. The invention relates to generating and/or identifying a compound that binds to and modulates (inhibits or enhances) the function of a component of a cell, thereby producing a phenotypic effect in the cell. Such a screen may involve identifying a biomolecule that 1) binds to, in vitro, a component of a cell that has been isolated from other constituents of the cell and that 2) causes, in vivo, as seen in an assay upon intracellular expression of the biomolecule, a phenotypic effect in the cell which is the usual producer and host of the target cell component. In an assay demonstrating characteristic 2) above, intracellular production of the biomolecule can be in cells grown in culture or in cells introduced into an animal. Further methods within these procedures are those methods comprising an assay for a phenotypic effect in the cell upon intracellular production of the biomolecule, either in cells in culture or in cells that have been introduced into one or more animals, and an assay to identify one or more compounds that behave as competitors of the biomolecule in an assay of binding to the target cell component. The target cell component in this embodiment and in other embodiments not limited to pathogens can be one that is found in mammalian cells, especially cells of a type found to cause or contribute to disease or the symptoms of disease (e.g., cells of tumors or cells of other types of hyperproliferative disorders).

1.7.1. Process for identifying one or more compounds that produce a phenotypic effect on a cell

One procedure envisioned in the invention is a process for identifying one or more compounds that produce a phenotypic effect on a cell. The process is at the same time a method for target validation. The process is characterized by identifying a biomolecule which binds an isolated target cell component, constructing cells comprising the target cell component and further comprising a gene encoding the biomolecular binder which can be expressed to produce the biomolecular binder, testing the constructed cells for their ability to produce, upon expression of the gene encoding the biomolecular binder, a phenotypic effect in the cells (e.g., inhibition of growth), wherein the test of the constructed cells can be a test of the cells in culture or a test of the cells after introducing them into host animals, or both, and further, identifying, for a biomolecular binder that caused the phenotypic effect, one or more compounds that compete with the biomolecular binder for binding to the target cell component.

A test of the constructed cells after introducing them into host animals is especially well-suited to assessing whether a biomolecular binder can produce a particular phenotype by the expression (regulatable by the researcher) of a gene encoding the biomolecular binder. In this method, cells are constructed which have a gene encoding the biomolecular binder, and wherein the biomolecular binder can be produced by regulation of expression of the gene. The constructed cells are introduced into a set of animals. Expression of the gene encoding the biomolecular binder is regulated in one group of the animals (test animals) such that the biomolecular binder is produced. In another group of animals, the gene encoding the biomolecular binder is regulated such that the biomolecular binder is not produced (control animals). The cells in the two groups of animals are monitored for a phenotypic change (for example, a change in growth rate). If the phenotypic change is observed in cells in the test animals and not in the cells in the control animals, or to a lesser extent in the control animals, then the biomolecular binder has been proven to be effective in binding to its target cell component under in vivo conditions.

A further embodiment of the invention is a method for determining whether a target cell component of a particular cell type (a "first cell") is essential to producing a phenotypic effect on the first cell, the method having the steps:
isolating the target component of the first cell; identifying a biomolecular binder of the isolated target component of the first cell; constructing a second type of cells ("second cell") comprising the target component and a regulable, exogenous gene encoding the biomolecular binder; and testing the second cell in culture for an altered phenotypic effect, upon production of the biomolecular binder in the second cell; whereby, if the second cell shows the altered phenotypic effect upon production of the biomolecular binder, then the target component of the first cell is essential to producing the phenotypic effect on the first cell. The target cell component in this embodiment and in other embodiments not limited to pathogens can be one that is found in mammalian cells, especially cells of a type found to cause or contribute to disease or the symptoms of disease (e.g., cells of tumors or cells of other types of hyperproliferative disorders).

1.7.3. Identifying a biomolecular inhibitor of growth of pathogen cells

One embodiment of the invention is a method for identifying a biomolecular inhibitor of growth of pathogen cells by using cell culture techniques, comprising contacting one or more types of biomolecules with isolated target cell component of the pathogen, applying a means of detecting bound complexes of biomolecules and target cell component, whereby, if the bound complexes are detected, one or more types of biomolecules have been identified as a biomolecular binder of the target cell component, constructing a pathogen strain having a regulatable gene encoding the biomolecular binder, regulating expression of the gene encoding the biomolecular binder to express the gene; and monitoring growth of the pathogen cells in culture relative to suitable control cells, whereby, if growth of the pathogen cells is decreased compared to growth of suitable control cells, then the biomolecule is a biomolecular inhibitor of growth of the pathogen cells.

1.7.4. Identifying compounds that inhibit infection of a mammal by a pathogen

A further embodiment of the invention is a method, employing an animal test, for identifying one or more compounds that inhibit infection of a mammal by a pathogen by binding to a target cell component, comprising constructing a pathogen comprising a regulable gene encoding a biomolecule which binds to the target cell component, infecting test animals with the pathogen, regulating expression of the regulable gene to produce the biomolecule, monitoring the test animals and suitable control animals for signs of infection, wherein observing fewer or less severe signs of infection in the test animals than in suitable control animals indicates that the biomolecule is a biomolecular inhibitor of infection, and identifying one or more compounds that compete with the biomolecular inhibitor of growth for binding to the target cell component (as by employing a competitive binding assay), then the compound inhibits infection of a mammal by a pathogen by binding to a target.

The competitive binding assay to identify binding analogs of biomolecular binders, which have been proven to bind to their targets in an intracellular test of binding, can be applied to any target for which a biomolecular binder has been identified, including targets whose function is unknown or targets for which other types of assays are not easily developed and performed. Therefore, the method of the invention offers the advantage of decreasing assay development time when using a gene product of known function as a target cell component and the advantage of bypassing the major hurdle of gene function identification when using a gene product of unknown function as a target cell component.

Other embodiments of the invention are cells comprising a biomolecule and a target cell component, wherein the biomolecule is produced by expression of a regulable gene, and wherein the biomolecule modulates function of the target cell component, thereby causing a phenotypic change in the cells. Yet other embodiments are cells comprising a biomolecule and a target cell component, wherein the biomolecule is a biomolecular binder of the target cell component, and is encoded by a regulatable gene. The cells can include mammalian cells or cells of a pathogen, for instance, and the phenotypic change can be a change in growth rate.

The pathogen can be a species of bacteria, yeast, fungus, or parasite, for example.

1.7.5. Intracellular validation of a biomolecule

Described herein are methods that result in the identification of compounds that cause a phenotypic effect on a cell. The general steps described herein to find a compound for drug development can be thought of as these: (1) identifying a biomolecule that can bind to an isolated target cell component in vitro, (2) confirming that the biomolecule, when produced in cells with the target cell component, can cause a desired phenotypic effect and (3) identifying, by an in vitro screening method, for example, compounds that compete with the biomolecule for binding to the target cell component. Central to these methods is general step (2) above, intracellular validation of a biomolecule comprising one or more steps that determine whether a biomolecule can cause a phenotypic effect on a cell, when the biomolecule is produced by the expression (which can be regulatable) of a gene in the cell. As used in general step (2), a biomolecule is a gene product (e.g., polypeptide, RNA, peptide or RNA oligonucleotide) of an exogenous gene -- a gene which has been introduced in the course of construction of the cell.

Biomolecules that bind to and alter the function of a candidate target are identified by various in vitro methods. Upon production of the biomolecule within a cell either in vitro or within an animal model system, the biomolecule binds to a specific site on the target, alters its intracellular function, and hence produces a phenotypic change (e.g. cessation of growth, cell death). When the biomolecule is produced in engineered pathogen cells in an animal model of infection, cessation of growth or death of the engineered pathogen cells leads to the clearing of infection and animal survival, demonstrating the importance of the target in infection and thereby validating the target.

A further embodiment of this invention provides for identifying a biomolecule that produces a phenotypic effect on a cell (wherein the cell can be, for instance, a pathogen cell or a mammalian cell) and (2) simultaneous intracellular target validation (see reference: Patents??).

1.7.6. Methods for identifying compounds that inhibit the growth of cells having a target cell component

The invention includes methods for identifying compounds that inhibit the growth of cells having a target cell component. The target cell component can first be identified as essential to the growth of the cells in culture and/or under conditions in which it is desired that the growth of the cells be inhibited. These methods can be applied, for example, to various types of cells that undergo abnormal or undesirable proliferation, including cells of neoplasms (tumors or growths, either benign or malignant) which, as known in the art, can originate from a variety of different cell types. Such cells can be referred to, for example, as being from adenomas, carcinomas, lymphomas or leukemias. The method can also be applied to cells that proliferate abnormally in certain other diseases, such as arthritis, psoriasis or autoimmune diseases.

If intracellular expression of the biomolecular binder inhibits the function of a target essential for growth (presumably by binding to the target at a biologically relevant site) cells monitored in step (2) will exhibit a slow growth or no growth phenotype. Targets found to be essential for growth by these methods are validated starting points for drug discovery, and can be incorporated into assays to identify more stable compounds that bind to the same site on the target as the biomolecule. Where the cells are pathogen cells and the desired phenotypic change to be monitored is inhibition of growth, the invention provides a procedure to examine the activity of target (pathogen) cell components in an animal infection model.

1.7.7. Study as a target cell component a gene product of a particular cell type

In the course of this method, it may be decided to study as a target cell component a gene product of a particular cell type (e.g., a type of pathogenic bacteria), wherein the target cell component is already known as being encoded by a characterized gene, as a potential target for a modulator to be identified. In this case, the target cell component can be isolated directly from the cell type of interest, assuming suitable culture methods are available to grow a sufficient number of cells, using methods appropriate to the type of cell component to be isolated (e.g., protein purification methods such as differential precipitation, ion exchange chromatography, gel chromatography, affinity chromatography, HPLC.

1.7.8. Target cell component can be produced recombinantly

Alternatively, the target cell component can be produced recombinantly, which requires that the gene encoding the target cell component be isolated from the cell type of interest. This can be done by any number of methods, for example known methods such as PCR, using template DNA isolated from the pathogen or a DNA library produced from the pathogen DNA, and using primers based on known sequences or combinations of known and unknown sequences within or external to the chosen gene. See, for example, methods described in "The Polymerase Chain Reaction," Chapter 15 of Current Protocols in Molecular Biology, (Ausubel, F.M. et al., eds), John Wiley & Sons, New York, 1998. Other methods include cloning a gene from a DNA library (e.g., a cDNA library from a eucaryotic pathogen) into a vector (e.g., plasmid, phage, phagemid, virus, etc.) and applying a means of selection or screening, to clones resulting from a transformation of vectors (including a population of vectors now having inserted genes) into appropriate host cells. The screening method can take advantage of properties given to the host cells by the expression of the inserted chosen gene (e.g., detection of the gene product by antibodies directed against it, detection of an enzymatic activity of the gene product), or can detect the presence of the gene itself (for instance, by methods employing nucleic acid hybridization). For methods of cloning genes in E. coli, which also may be applicable to cloning in other bacterial species, see, for example, "Escherichia coli, Plasmids and Bacteriophages," Chapter I of Current Protocols in Molecular Biology, (Ausubel, F.M. et al., eds), John Wiley & Sons, New York, 1998. For methods applicable to cloning genes of eukaryotic origin, see Chapter 5 ("Construction of Recombinant DNA Libraries"), Chapter 9 ("Introduction of DNA Into Mammalian Cells") and Chapter 6 ("Screening of Recombinant DNA Libraries") of Current Protocols in Molecular Biology, (Ausubel, F.M. et al., eds), John Wiley & Sons, New York, 1998.

Target proteins can be expressed with E. coli or other prokaryotic gene expression systems, or in eukaryotic gene expression systems. Since many eukaryotic proteins carry unique modifications that are required for their activities, e.g. glycosylation and methylation, protein expression can in some cases be better carried out in eukaryotic systems, such as yeast, insect, or mammalian cells that can perform these modifications. Examples of these expression systems have been reviewed in the following literature: Methods in Enzymology, Volume 185, eds D.V. Goeddel, Academic Press, San Diego, 1990; Geisse et al, Protein Expression and Purification 8:271-282, 1996; Simonsen and McGrogan, Biologicals 22: 85-94; Jones and Morikawa, Current Opinions in Biotechnologies 7: 512-516, 1996; Possee, Current Opinions in Biotechnologies 8:569-572.

Where a gene encoding a chosen target cell component has not been isolated previously, but is thought to exist because homologs of the gene product are known in other species, the gene can be identified and cloned by a method such as that used in Shiba et al., US 5,759,833, Shiba et al., US 5,629,188, Martinis et al., US 5,656,470 and Sassanfar et al., US 5,756,327. The teachings of these four patents are incorporated herein by reference in their entirety.

1.7.9. Method should be used with target cell components which have not been previously isolated or characterized and whose functions are unknown

It is an advantage of the target validation method that it can be used with target cell components which have not been previously isolated or characterized and whose functions are unknown. In this case, a segment of DNA containing an open reading frame (ORF; a cDNA can also be used, as appropriate to a eukaryotic cell) which has been isolated from a cell of a type that is to be an object of drug action (e.g., tumor cell, pathogen cell) can be cloned into a vector, and the target gene product of the ORF can be produced in host cells harboring the vector. The gene product can be purified and further studied in a manner similar to that of a gene product that has been previously isolated and characterized.

In some cases, the open reading frame (in some cases, cDNA) can be isolated from a source of DNA of the cells of interest (genomic DNA or a library, as appropriate), and inserted into a fusion protein or fusion polypeptide construct. This construct can be a vector comprising a nucleic acid sequence which provides a control region (e.g., promoter, ribosome binding site) and a region which encodes a peptide or polypeptide portion of the fusion polypeptide wherein the polypeptide encoded by the fusion vector endows the fusion polypeptide with one or more properties that allow for the purification of the fusion polypeptide. For example, the vector can be one from the pGEX series of plasmids (Pharmacia) designed to produce fusions with glutathione S-transferase.

1.7.10. Host cells

The isolated DNA having an open reading frame, whether encoding a known or an as yet unidentified gene product, when inserted into an expression construct, can be expressed to produce the target cell component in host cells. Host cells can be, for example, Gram-negative or Gram-positive bacterial cells such as Escherichia coli or Bacillus subtilis, respectively, or yeast cells such as Saccharomyces cerevisiae, Schizosaccharomyces pombe or Pichia pastoris. It is preferable that the target cell component to be used in target validation studies be produced in a host that is genetically related to the pathogen from which the gene encoding it was isolated. For example, for a Gram-negative bacterial pathogen, an E. coli host is preferred over a Pichia pastoris host. The target cell component so produced can then be isolated from the host cells. Many protein purification methods are known that separate proteins on the basis of, for instance, size, charge, or affinity for a binding partner (e.g., for an enzyme, a binding partner can be a substrate or substrate analog), and these methods can be combined in a sequence of steps by persons of skill in the art to produce an effective purification scheme. For methods to manipulate RNA, see, for example, Chapter 4 in Current Protocols in Molecular Biology (Ausubel, F.M. et al., eds), John Wiley & Sons, New York, 1998.

An isolated cell component or a fusion protein comprising the cell component can be used in a test to identify one or more biomolecular binders of the isolated product (general step (1)). A biomolecular binder of a target cell component can be identified by in vitro assays that test for the formation of complexes of target and biomolecular binder noncovalently, bound to each other. For example, the isolated target can be contacted with one or more types of biomolecules under conditions conducive to binding, the unbound biomolecules can be removed from the targets, and a means of detecting bound complexes of biomolecules and targets can be applied. The detection of the bound complexes can be facilitated by having either the potential biomolecular binders or the target labeled or tagged with an adduct that allows detection or separation (e. g., radioactive isotope or fluorescent label; streptavidin, avidin or biotin affinity label).

Alternatively, both the potential biomolecular binders and the target can be differentially labeled. For examples of such methods see, e.g., WO 98/19162.

1.7.11. Biomolecules to be tested and means for detection

The biomolecules to be tested for binding to a target can be from a library of candidate biomolecular binders, (e.g., a peptide or oligonucleotide library). For example, a peptide library can be displayed on the coat protein of a phage (see, for examples of the use of genetic packages such as phage display libraries, Koivunen, E. et al., J Biol. Chem. 268:20205-20210 (1993)). The biomolecules can be detected by means of a chemical tag or label attached to or integrated into the biomolecules before they are screened for binding properties. For example, the label can be a radioisotope, a biotin tag, or a fluorescent label. Those molecules that are found to bind to the target molecule can be called biomolecular binders.

1.7.12. Fusion proteins

An isolated target cell component, an antigenically similar portion thereof, or a suitable fusion protein comprising all of or a portion of or the entire target can be used in a method to select and identify biomolecules which bind specifically to the target. Where the target cell component comprises a protein, fusion proteins comprising all of, or a portion of, the target linked to a second moiety not occurring in the target as found in nature, can be prepared for use in another embodiment of the method. Suitable fusion proteins for this purpose include those in which the second moiety comprises an affinity ligand (e.g., an enzyme, antigen, epitope). The fusion proteins can be produced by the insertion of a gene encoding a target or a suitable portion of such gene into a suitable expression vector, which encodes an affinity ligand (e.g., pGEX-4T-2 and pET- 15b, encoding glutathione S- transferase and His-Tag affinity ligands, respectively). The expression vector can be introduced into a suitable host cell for expression. Host cells are lysed and the lysate, containing fusion protein, can be bound to a suitable affinity matrix by contacting the lysate with an affinity matrix under conditions sufficient for binding of the affinity ligand portion of the fusion protein to the affinity matrix.

1.7.12.1. Fusion protein can be immobilized

In one embodiment, the fusion protein can be immobilized on a suitable affinity matrix under conditions sufficient to bind the affinity ligand portion of the fusion protein to the matrix, and is contacted with one or more candidate biomolecules (e.g., a mixture of peptides) to be tested as biomolecular binders, under conditions suitable for binding of the biomolecules to the target portion of the bound fusion protein. Next, the affinity matrix with bound fusion protein can be washed with a suitable wash buffer to remove unbound biomolecules and non- specifically bound biomolecules. Biomolecules which remain bound can be released by contacting the affinity matrix with fusion protein bound thereto with a suitable elution buffer. Wash buffer can be formulated to permit binding of the fusion protein to the affinity matrix, without significantly disrupting binding of specifically bound biomolecules. In this aspect, elution buffer can be formulated to permit retention of the fusion protein by the affinity matrix, but can be formulated to interfere with binding of the test biomolecule(s) to the target portion of the fusion protein. For example, a change in the ionic strength or pH of the elution buffer can lead to release of biomolecules, or the elution buffer can comprise a release component or components designed to disrupt binding of biomolecules to the target portion of the fusion protein.

Immobilization can be performed prior to, simultaneous with, or after contacting, the fusion protein with biomolecule, as appropriate. Various permutations of the method are possible, depending upon factors such as the biomolecules tested, the affinity matrix-ligand pair selected, and elution buffer formulation. For example, after the wash step, fusion protein with biomolecules bound thereto can be eluted from the affinity matrix with a suitable elution buffer (a matrix elution buffer, such as glutathione for a GST fusion). Where the fusion protein comprises a cleavable linker, such as a thrombin cleavage site, cleavage from the affinity ligand can release a portion of the fusion with the biomolecules bound thereto. Bound biomolecule can then be released from the fusion protein or its cleavage product by an appropriate method, such as extraction.

1.7.12. Various methods to identify biomolecular binders

One or more candidate biomolecular binders can be tested simultaneously. Where a mixture of biomolecules is tested, the biomolecules selected by the foregoing processes can be separated (as appropriate) and identified by suitable methods (e.g., PCR, sequencing, chromatography). Large libraries of biomolecules (e.g., peptides, RNA oligonucleotides) produced by combinatorial chemical synthesis or other methods can be tested (see e. a., Ohlmeyer, M.H.J. et al., Proc. Natl. Acad. Sci. USA 90:10922-10926 (1993) and DeWitt, S.H. et al., Proc. Natl. Acad. Sci. USA 90:6909-6913 (1993), relating to tagged compounds; see also Rutter, W.J. et al. U.S. Patent No. 5,010,175; Huebner, V.D. et al., U.S. Patent No. 5,182,366; and Geysen, H.M., U.S. Patent No. 4,833,092). Random sequence RNA libraries (see Ellington, A.D. et al., Nature 346:818-822 (1990); Bock, L.C. et al., Nature 355:584-566 (1992); and Szostak, J.W., Trends in Biochem. Sci. 17:89-93 (March, 1992)) can also be screened according to the present method to select RNA molecules which bind to a target. Where biomolecules selected from a combinatorial library by the present method carry unique tags, identification of individual biomolecules by chromatographic methods is possible. Where biomolecules do not carry tags, chromatographic separation, followed by mass spectrometry to ascertain structure, can be used to identify individual biomolecules selected by the method, for example.

Other methods to identify biomolecular binders of a target cell component can be used. For example, the two-hybrid system or interaction trap is an in vivo system that can be used to identify polypeptides, peptides or proteins (candidate biomolecular binders) that bind to a target protein. In this system, both candidate biomolecular binders and target cell component proteins are produced as fusion proteins. The two-hybrid system and variations on it have been described (US 5,283,173 and US 5,468,614; Golemis, E.A. et al., pages 20.1.1-20.1.35 In Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., John

Wiley and Sons, containing supplements up through Supplement 40, 1997; two-hybrid systems available from Clontech, Palo Alto, CA).

Once one or more biomolecular binders of a cell component have been identified, further steps can be combined with those taken to identify the biomolecular binder, to identify those biomolecular binders that produce a phenotypic effect on a cell (where "a cell" can mean cells of a cell strain or cell line).

Thus, a method for identifying a biomolecule that produces a phenotypic effect on a first cell can comprise the steps of identifying a biomolecular binder of an isolated target cell component of the first cell, constructing a second cell comprising the target cell component and a regulable exogenous gene encoding the biomolecular binder, and testing the second cell for the phenotypic effect, upon production of the biomolecular binder in the second cell, where the second cell can be maintained in culture or introduced into an experimental animal. If the second cell shows the phenotypic effect upon intracellular production of the biomolecular binder, then a biomolecule that produces a phenotypic effect on the first cell has been identified. Testing the second cell is general step (2) of the invention, as the three general steps were outlined above.

1.7.13. Host cells: Engineered to control expression

Host cells (also, "second cells" in the terminology used above) of the cell type (e.g., species of pathogenic bacteria) the target was isolated from (or the gene encoding the target was originally isolated from, if the target is produced by recombinant methods), can be engineered to harbor a gene that can regulably express the biomolecular binder (e.g., under an inducible or repressible promoter). The ability to regulate the expression of the biomolecular binder is desirable because constitutive expression of the biomolecular binder could be lethal to the cell.

Therefore, inducible or regulated expression gives the researcher the ability to control if and when the biomolecular binder is expressed. The gene expressing the biomolecular binder can be present in one or more copies, either on an extra chromosomal structure, such as on a single or multicopy plasmid, or integrated into the host cell genome. Plasmids that provide an inducible gene expression system in pathogenic organisms can be used. For example, plasmids allowing tetracycline-inducible expression of a gene in Staphylococcus aureus have been developed.

1.7.14. Genes for expression

For intracellular expression of a biomolecule to be tested for its phenotypic effect in a eukaryotic cell (e.g., mammalian cell), the genes for expression can be carried on plasmid-based or virus-based vectors, or on a linear piece of DNA or RNA. For examples of expression vectors, see Hosfield and Lu, Biotechniques: 306-309, 1998; Stephens and Cockett, Nucleic Acid Research 17:7110, 1989; Wohlgemuth et al, Gene Therapy, 3:503-512, 1996; Ramirez-Solis et al, Gene 87:291-294, 1990, Dirks et al, Gene 149:387-388, 1994; Chenaalvala et al. Current Opinion in Biotechnologies 2:718-722, 1991; Methods in Enzymology, Volume 185, (D.V. Goeddel, ed.) Academic Press, San Diego, 1990. The genetic material can be introduced into cells using a variety of techniques, including whole cell or protoplast transformation, electroporation, calcium phosphate-DNA precipitation or DEAE- Dextran transfection, liposome mediated DNA or RNA transfer, or transduction with recombinant viral or retroviral vectors. Expression of the gene can be constitutive (e.g., ADHI promoter for expression in S. cerevisiae (Bennetzen, J.L. and Hall, B.D., J Biol. Chem 257:3026-3031 (1982)), or CMV immediate early promoter and RSV LTR for mammalian expression) or inducible, as the inducible GAL I promoter in yeast (Davis, L.I. and Fink, G.R., Cell 61:965-978 (1990)). A variety of inducible systems can be utilized, for example, E. coli Lac repressor/operator system and Tn10 Tet repressor/operator systems have been engineered to govern regulated expression in organisms from bacterial to mammalian cells. Regulated gene expression can also be achieved by activation. For example, gene expression governed by HIV LTR can be activated by HIV or SIV Tat proteins in human cells; GAL4 promoter can be activated by galactose in a nonglucose-containing medium. The location of the biomolecule binder genes can be extra chromosomal or chromosomally integrated. The chromosome integration can be mediated through homologous or nonhomologous recombinations.

For proper localization in the cells, it maybe desirable to tag the biomolecule binders with certain peptide signal sequences (for example, nuclear localization signal (NLS) sequences, mitochondria localization sequences). Secretion sequences have been well documented in the art.

1.7.15. Fused biomolecular binders

For presentation of the biomolecular binders in the intracellular system, they can be fused N-terminally, C-terminally, or internally in a carrier protein (if the biomolecular binder is a peptide), and can be fused (5', 3' or internally) in a carrier RNA or DNA molecule (if the biomolecular binder is a nucleic acid). The biomolecular binder can be presented with a protein or nucleic acid structural scaffold. Certain linkages (e.g., a 4-glycine linker for a peptide or a stretch of A's for an RNA can be inserted between the biomolecular binder and the carrier proteins or nucleic acids.

In such engineered cells, the effect of this biomolecular binder on the phenotype of the cells can be tested, as a manifestation of the binding (implying binding to a functionally relevant site, thus, an activator, or more likely, an inhibitory) effect of the biomolecular binder on the target used in an in vitro binding assay as described above. An intracellular test can not only determine which biomolecular binders have a phenotypic effect on the cells, but at the same time can assess whether the target in the cells is essential for maintaining the normal phenotype of the cells. For example, a culture of the engineered cells expressing a biomolecular binder can be divided into two aliquots. The first aliquot ("test" cells) can be treated in a suitable manner to regulate (e.g., induce or release repression of, as appropriate) the gene encoding the biomolecular binder, such that the biomolecular binder is produced in the cells. The second aliquot ("control" cells) can be left untreated so that the biomolecular binder is not produced in the cells. In a variation of this method of testing the effect of a biomolecular binder on the phenotype of the cells, a different strain of cells, not having a gene that can express the biomolecular binder, can be used as control cells. The phenotype of the cells in each culture ("test" and "control" cells grown under the same conditions, other than the expression of the biomolecular binder), can then be monitored by a suitable means (e.g., enzymatic activity, monitoring, a product of a biosynthetic pathway, antibody to test for presence of cell surface antigen, etc.). Where the change in phenotype is a change in growth rate, the growth of the cells in each culture ("test" and "control" cells grown under the same conditions, other than the expression of the biomolecular binder), can be monitored by a suitable means (e.g., turbidity of liquid cultures, cell count, etc). If the extent of growth, or rate of growth of the test cells is less than the extent of growth or rate of growth of the control cells, then the biomolecular binder can be concluded to be an inhibitor of the growth of the cells, or a biomolecular inhibitor.

If the phenotype of the test cells is altered relative to that of the control cells, then the biomolecular binder can be concluded to be one that causes a phenotypic effect. In an optional additional test, isolated target cell component having a known function (e.g., an enzyme activity) can be tested for modulation of this known function in the presence of biomolecular binder under conditions conducive to binding of the biomolecular binder to the target cell component. Positive results in these tests should encourage the investigator to continue in the drug discovery process with efforts to find a more stable compound (than a peptide, polypeptide or RNA biomolecule) that mimics the binding properties of the biomolecular binder on the tested target cell component.

1.7.16. Engineering strain of cells

A further test can, again, employ an engineered strain of cells that comprise both the target cell component and one or more genes encoding a biomolecule tested to be a biomolecular binder of the target cell component. The cells of the cell strain can be tested in animals to see if regulable expression of the biomolecular binder in the engineered cells produces an observable or testable change in phenotype of the cells. Both the "in culture" test for the effect of intracellular expression of the biomolecular binder and the "in animal" test (described below) for the effect of intracellular expression of the biomolecular binder can be applied not only towards drug discovery in the categories of antimicrobials and anticancer agents, but also towards the discovery of therapeutic agents to treat inflammatory diseases, cardiovascular diseases, diseases associated with metabolic pathways, and diseases associated with the central nervous system, for example.

Where the engineered strain of cells is a strain of pathogen cells or tumor cells, the object of the test is to see whether production of the biomolecular binder in the engineered strain inhibits growth of these cells after their introduction into an animal by the engineered pathogen. Such a test can not only determine which biomolecular binders are inhibitors of growth of the cells, but at the same time can assess whether the target in the cells is essential for maintaining growth of the cells (infection, for a pathogenic organism) in a host mammal. Suitable animals for such an experiment are, for example, mammals such as mice, rats, rabbits, guinea pigs, dogs, pigs, and the like. Small mammals are preferred for reasons of convenience.

The engineered cells are introduced into one or more animals ("test" animals) and into one or more animals in a separate group ("control" animals) by a route appropriate to cause symptoms of systemic or local growth of the engineered cells.

The route of introduction may be, for example, by oral feeding, by inhalation, by subdermal, intramuscular, intravenous, or intraperitoneal injection as appropriate to the desired result.

After the cell strain has been introduced into the test and control animals, expression of the gene encoding the biomolecular binder is regulated to allow production of the biomolecular binder in the engineered pathogen cells. This can be achieved, for instance, by administering to the test animals a treatment appropriate to the regulation system built into the cells, to cause the gene encoding the biomolecular binder to be expressed. The same treatment is not administered to the control animals, but the conditions under which they are maintained are otherwise identical to those of the test animals. The treatment to express the gene encoding the biomolecular binder can be the administration of an inducer substance (where expression of the biomolecular binder or gene is under the control of an inducible promoter) or the functional removal of a repressor substance (where expression of the biomolecular binder gene is under the control of a repressible promoter).

After such treatment, the test and control animals can be monitored for a phenotypic effect in the introduced cells. Where the introduced cells are constructed pathogen cells, the animals can be monitored for signs of infection (as the simplest endpoint, death of the animal, but also e.g., lethargy, lack of grooming behavior, hunched posture, not eating, diarrhea or other discharges; bacterial titer in samples of blood or other cultured fluids or tissues). In the case of testing engineered tumor cells, the test and control animals can be monitored for the development of tumors or for other indicators of the proliferation of the introduced engineered cells. If the test animals are observed to exhibit less growth of the introduced cells than the control animals, then the biomolecule can be also called a biomolecular inhibitor of growth, or biomolecular inhibitor of infection, as appropriate, as it can be concluded that the expression in vivo of the biomolecular inhibitor is the cause of the relative reduction in growth of the introduced cells in the test animals.

1.7.17. In vitro assays

Further steps of the procedure involve in vitro assays to identify one or more compounds that have binding and activating or inhibitory properties that are similar to those of the biomolecules which have been found to have a phenotypic effect, such as inhibition of growth. That is, compounds that compete for binding to a target cell component with the biomolecule would then be structural analogs of the biomolecules. Assays to identify such compounds can take advantage of known methods to identify competing molecules in a binding assay. These steps comprise general step (3) of the method.

In one method to identify such compounds, a biomolecular inhibitor (or activator) can be contacted with the isolated target-cell component to allow binding, one or more compounds can be added to the milieu comprising the biomolecular inhibitor and the cell component under conditions that allow interaction and binding between the cell component and the biomolecular inhibitor, and any biomolecular inhibitor that is released from the cell component can be detected.

1.7.18. Fluorescence

One suitable system that allows the detection of released biomolecular inhibitor (or activator) is one in which fluorescence polarization of molecules in the milieu can be measured. The biomolecular inhibitor can have bound to it a fluorescent tag or label such as fluorescein or fluorescein attached to a linker.

Assays for inhibition of the binding of the biomolecular inhibitor to the cell component can be done in microtiter plates to conveniently test a set of compounds at the same time. In such assays, a majority of the fluorescently labeled biomolecular inhibitor must bind to the protein in the absence of competitor compound to allow for the detection of small changes in the bound versus free probe population when a compound which is a competitor with a biomolecular inhibitor is added (B.A. Lynch, et al., Analytical Biochemistry 247:77-82 (1997)). If a compound competes with the biomolecular inhibitor for a binding site on the target cell component, then fluorescently labeled biomolecular inhibitor is released from the target cell component, lowering the polarization measured in the milieu.

1.7.19. Radioactive isotope

In a further method for identifying one or more compounds that compete with a biomolecular inhibitor (or activator) for a binding site on a target cell component, the target cell component can be attached to a solid support, contacted with one or more compounds, and contacted with the biomolecular inhibitor. One or more washing steps can be employed to remove biomolecular inhibitor and compound not bound to the cell component. Either the biomolecular inhibitor bound to the target cell component or the compound bound to the target cell component can be measured. Detection of biomolecular inhibitor or compound bound to the cell compound can be facilitated by the use of a label on either molecule type, wherein the label can be, for instance, a radioactive isotope either incorporated into the molecule itself or attached as an adduct, streptavidin or biotin, a fluorescent label or a substrate for an enzyme that can produce from the substrate a colored or fluorescent product. An appropriate means of detection of the labeled biomolecular inhibitor or compound moiety of the biomolecular inhibitor- cell component complex or the compound-cell component complex can be applied. For example, a scintillation counter can be used to measure radioactivity. Radio labeled streptavidin or biotin can be allowed to bind to biotin or streptavidin, respectively, and the resulting complexes detected in a scintillation counter. Alkaline phosphatase conjugated to streptavidin can be added to a biotin-labeled biomolecular inhibitor or compound. Detection and quantitation of a biotin-labeled complex can then be by addition of pNPP substrate of alkaline phosphatase and detection by spectrophotometry, of a product which absorbs UV light at a wavelength of 405 nm. A fluorescent label can also be used, in which case detection of fluorescent complexes can be by a fluorometer. Models are available that can read multiple samples, as in a microtiter plate.

For example, in one type of assay, the method for identifying compounds comprises attaching the target cell component to a solid support, contacting the biomolecular inhibitor with the target cell component under conditions suitable for binding of the biomolecular inhibitor to the cell component, removing unbound biomolecular inhibitor from the solid support, contacting one or more compounds (e.g., a mixture of compounds) with the cell component under conditions suitable for binding of the biomolecular inhibitor to the cell component, and testing for unbound biomolecular inhibitor released from the cell component, whereby if unbound biomolecular inhibitor is detected, one or more compounds that displace or compete with the biomolecular inhibitor for a particular site on the target cell component have been identified.

Other methods for identifying compounds that are competitive binders with the biomolecule for a target can employ adaptations of fluoresence polarization methods. See, for instance, Anal. Biochem. 253(2):210-218 (1997), Anal. Biochem. 249(l):29-36 (1997), BioTechniques 17(3):585-589 (1994) and Nature 373:254-256 (1995).

Those compounds that bind competitively to the target cell component can be considered to be drug candidates. Further appropriate testing can confirm that those compounds which bind competitively with biomolecular inhibitors (or activators) possess the same activity as seen in an intracellular test of the effect of the biomolecular inhibitor or activator upon the phenotype of cells. Derivatives of these compounds having modifications to confer improved solubility, stability, etc., can also be tested for a desired phenotypic effect.

1.7.20. Combining steps

Combining steps for testing the phenotypic effects of a biomolecule, as can be produced in an intracellular test, with steps for identifying compounds that compete with the biomolecule for sites on a target cell component, yields a method for identifying a compound which is a functional analog of a biomolecule which produces a phenotypic effect on a cell. These steps can be to test, for the phenotypic effect, either in culture or in an animal model, or in both, a cell which produces a biomolecule by regulable expression of an exogenous gene in the cell, and to identify, if the biomolecule caused the phenotypic effect, one or more compounds that compete with the biomolecule for binding to a target cell component. If a compound is found to compete with the biomolecule for binding to the target cell component, then the compound is a functional analog of a biomolecule which produces a phenotypic effect on the cell. Such a functional analog can cause qualitatively a similar effect on the cell, but to a similar degree, lesser degree or greater degree than the biomolecule.

1.7.21. Method for determining whether a target component of a cell is essential to producing a phenotypic effect on the cell

A further embodiment of the invention combining general steps (1) and (2) is a method for determining whether a target component of a cell is essential to producing a phenotypic effect on the cell, comprising isolating the target component from the cell, identifying a biomolecular binder of the isolated target component of the cell, constructing a second cell comprising the target component and a regulable, exogenous gene encoding the biomolecular binder, and testing the second cell in culture for an altered phenotypic effect, upon production of the biomolecular binder in the second cell, whereby, if the second cell shows the altered phenotypic effect upon production of the bimolecular binder, then the target component of the first cell is essential to producing the phenotypic effect on the first cell.

1.7.22. Inhibit the proliferation of the cells

The methods described herein are well suited to the identification of compounds that can inhibit the proliferation of the cells of infectious agents such as bacteria, fungi and the like. In addition, a procedure such as the one outlined below can be used in the identification of compounds to inhibit the proliferation of cancer cells. The two procedures described below further illustrate the use of the methods described herein and would provide proof of principle of these methods with a known target for anticancer therapy.

Mammalian dihydrofolate reductase (DHFR) is a proven target for anticancer therapy. Methotrexate (MTX) is one of many existing drugs that inhibit DHFR. It is widely used for anticancer chemotherapy.

NIH 3T3 is a mouse fibroblast cell line that is able to develop spontaneous transformed cells when cultured in low concentration (2%) of calf serum in molecular, cellular and developmental biology medium 402 (MCDB) (M. Chow and H. Rubin, Proc. Natl. Acad. Sci. USA 95(8):4550-4555 (1998)). The transformed cells, which can be selectively inhibited by MTX (Chow and Rubin), are isolated.

Both the normal and transformed NIH3T3 cells are transfected with pTet- On plasmid (Clontech; Palo Alto, CA). Stable cell lines that express high levels of reverse tetracycline-control led activator (rtTA) are isolated and characterized for their normal or transformed phenotype (Chow and Rubin).

The DHFR gene (Genbank Accession # L26316) from the NIH 3T3 cell line is amplified by reverse transcription-PCR (RT-PCR) using poly A' RNA isolated from NIH 3T3 cells (Sambrook, J. et al., Molecular Cloning: A Laboratory Manual, 2nd edition, Cold Spring Harbor Laboratory Press, 1989). Active DHFR is expressed using the BacPAK Baculovirus Expression System (Clontech) or other appropriate systems. The expressed DHFR is purified and biotinylated and subjected to peptide binder identification as exemplified for bacterial proteins. The identified peptides are biochemically characterized for in vitro inhibition of DHFR activity. Peptides that inhibit DHFR are identified. A nucleic acid encoding each peptide can be cloned into a vector such as pGEX-4T2 (Pharmacia) to yield a vector which encodes a fusion polypeptide having the peptide fused to the N- terminus of GST. This can also be done by PCR amplification as exemplified herein for the peptide Pro-3. The fusion genes are cloned into plasmid pTRE (Clontech) for regulated expression. The constructed plasmid or the vector is cotransfected with pTK-Hyg into the stable NIH 3T3 cell line that expresses rtTA. The resulting cell lines, termed 3T3N-VITA (normal 3T3 cells that express rtTA and the DHFR inhibitory peptides), 3T3T-VITA (transformed 3T3 cells that express rtTA and the DHFR inhibitory peptides), or 3T3T-VITA control (transformed 3T3 cells that express rtTA and GST), are characterized for their normal or transformed phenotype (loss of contact inhibition, change in morphology, immortalization, etc.). 10^2-10^1 of 3T3T-VITA or 3T3T-VITA control cells are mixed with 10^5 3T3N-VITA and are grown in MCD 402 medium with 10% calf serum at 37'C for three days. Tetracycline is added to the medium to a final concentration of 0 to 1 ug/ml. In a control, 200 nM of MTX is added. The cultures are incubated for an additional eight days, and the number of foci formed are counted as described by M. Chow and H. Rubin, Proc. Natl. Acad Sci. USA 95(8):4550-4555 (1998). Peptides that specifically inhibit foci formation of 3T3 transformed cells are identified.

A murine model of fibroblastoma (Kogerman, P. et al., Oncogene (12):1407-1416 (1997)) is used for evaluating the DHFR/peptide combination for identification of compounds for cancer therapy. Various amounts of 3T3T- VITA or 3T3T-VITA control cells (10^3, 10^4, 10^5, 10^6 cells) are injected subcutaneously into 5 groups (10 in each group) of athymic nude mice (4-6 weeks old, 18-22 g) to determine the minimal dose needed for development of fibroblastomas in all of the tested animals. Upon determination of the minimal tumorigenic dose, 6 groups of athymic nude mice (10 each) are injected subcutaneously (s.c.) with the minimal tumorigenic dose for 3T3T-VITA or 3T3T-VITA control cells to develop fibroblastoma. One week after injection, group I mice start receiving MTX s.c. at 2 mg/kg/day as positive control, group 2 to 5 start receiving 1, 2, 5, or 10 mg/kg/day of tetracycline, group 6 start receiving saline (vehicle) as control. Five weeks after the introduction of cells, all of the mice are sacrificed and tumors are removed from them. Tumor mass is measured and compared among the groups.

An effective peptide identified by these in vivo experiments can be used for screening libraries of compounds to identify those compounds that competitively bind to DHFR. One mechanism of tumorigenesis is overexpression of proto-oncogenes such as Ha-ras (Reviewed by Suarez, H.G.,Anticancer Research 9(5):1331-1343 (1989)).

Compounds that inhibit the activities of the products of such proto- oncogenes can be used for cancer chemotherapy. What follows is a further illustration of the methods described herein, as applied to mammalian cells.

Transgenic mice that overexpress human Ha-ras have been produced. Such transgenic mice develop salivary and/or mammary adenocarcinomas (Nielsen, L.L. et al, In Vivo 8(5):1331-1343 (1994)). Secondary transgenic mice that express rtTA can be generated using the pTet-On plasmid from Clontech.

Human Ha-ras open reading frame cDNA (Genbank Accession #GO0277) is amplified by RT-PCR using polyA- RNA isolated from human mammary gland or other tissues. Active Ha-ras is expressed using the BacPAK Baculovirus Expression System (Clontech) or other appropriate systems. The expressed Ha-ras is purified and biotinylated and subjected to peptide binder identification as exemplified herein for bacterial proteins as target cell components. The identified peptides are biochemically characterized for in vitro inhibition of Ha-ras GTPase activity.

Peptides that inhibit Ha-ras are cloned into plasmid pTPE (Clontech) for regulated expression as an N-terminal fusion of GST. Such constructs are used to generate tertiary transgenic mice using the secondary transgenic mice. Transgenic mice that are able to overexpress peptide genes are identified by Northern and Western analysis. Control mice that express GST are also identified.

Various doses of tetracycline are administered to the tertiary transgenic mice by s.c. or i.p. injection before or after tumor onset. Prevention or regression of tumors resulting from expression of the peptide genes are analyzed as described above for murine fibroblastoma.

Peptides found to be effective in in vivo experiments will be used to screen compounds that inhibit human Ha-ras activity for cancer therapy.

1.7.23. Disease targets

The method of the invention can be applied more generally to mammalian diseases caused by: (1) loss or gain of protein function, (2) over- expression or loss of regulation of protein activity. In each case the starting point is the identification of a putative protein target or metabolic pathway involved in the disease. The protocol can sometimes vary with the disease indication, depending on the availability of cell culture and animal model systems to study the disease. In all cases the process can deliver a validated target and assay combination to support the initiation of drug discovery.

Appropriate disease indications include, but are not limited to, Alzheimer's, arthritis, cancer, cardiovascular diseases, central nervous system disorders, diabetes, depression, hypertension, inflammation, obesity and pain.

Appropriate protein targets putatively linked to disease indications include, but are not limited to (1) the leptin protein, putatively linked to obesity and diabetes; (2) a mitogen- activated protein kinase putatively linked to arthritis, osteoporosis and atherosclerosis; (3) the interleukin-1 beta converting protein putatively linked to arthritis, asthma and inflammation; (4) the caspase proteins putatively linked to neurodegenerative diseases such as Alzheimer's, Parkinson's and stroke, and (5) the tumor necrosis factor protein putatively linked to obesity and diabetes. Appropriate protein targets include also, but are not limited to, enzymes catalyzing the following types of reactions: (1) oxido-reductases, (2) transferases, (3) hydrolases, (4) lyases, (5) isomerases, and (6) ligases.

The arachidonic acid pathway constitutes one of the main mechanisms for the production of pain and inflammation. The pathway produces different classes of end products, including the prostaglandins, thromboxane and leukotrienes.

Prostaglandins, an end product of cyclooxygenase metabolism, modulate immune function, mediate vascular phases of inflammation and are potent vasodilators. The major therapeutic action of aspirin and other non-steroidal anti - inflammatory drugs (NSAIDs) is proposed to be inhibition of the enzyme cyclooxygenase (COX). Anti- inflammatory potencies of different NSAIDs have been shown to be proportional to their action as COX inhibitors. It has also been shown that COX inhibition produces toxic side effects such as erosive gastritis and renal toxicity. The knowledge base regarding the toxic side effects of COX inhibitors has been gained through years of monitoring human therapies and human suffering. Two kinds of COX enzymes are now known to exist, with inhibition of COX 1 related to toxicity, and inhibition of COX2 related to reduction of inflammation. Thus, selective COX2 inhibition is a desirable characteristic of new anti-inflammatory drugs. The method of the invention can provide a route from identification of potential drug targets to validating these targets (for example, COX1 and COX2) as playing a role in disease (pain and inflammation) to an examination of the phenotype for the inhibition of one or both target isozymes without human suffering. Importantly, this information can be collected in vivo.

As an alternative strategy, the method of the invention can be used to define the phenotype of "genes of unknown function" obtained from various human genome sequencing projects or to assess the phenotype resulting, from inhibition of one isozyme subtype or one member of a family of related protein targets.

1.5. Definitions

Target: (also, "target component of a cell," or "target cell component") a constituent of a cell which contributes to and is necessary for the production or maintenance of a phenotype of the cell in which it is found. A target can be a single type of molecule or can be a complex of molecules. A target can be the product of a single gene, but can also be a complex comprising more than one gene product (for example, an enzyme comprising alpha and beta subunits, mRNA, tRNA, ribosomal RNA or a ribonucleoprotein particle such as a snRNP). Targets can be the product of a characterized gene (gene of known function) or the product of an uncharacterized gene (gene of unknown function).

Target Validation: the process of determining whether a target is essential to the maintenance of a phenotype of the cell type in which the target normally occurs.
For example, for pathogenic bacteria, researchers developing antimicrobials want to know if a compound which is potentially an antimicrobial agent not only binds to a target in vitro, but also binds to, and modulates the function of, a target in the bacteria in vivo, and especially under the conditions in which the bacteria are producing an infection -- those conditions under which the antimicrobial agent must work to inhibit bacterial growth in an infected animal or human. If such compounds can be found that bind to a target in vitro and alter the target's function in cells resulting in an altered phenotype, as found by testing cells in culture and/or as found by testing cells in an animal, then the target is validated.

Phenotypic Effect: a change in an observable characteristic of a cell which can include, e.g., growth rate, level or activity of an enzyme produced by the cell, sensitivity to various agents, antigenic characteristics, and level of various metabolites of the cell. A phenotypic effect can be a change away from wild type (normal) phenotype, or can be a change towards wild type phenotype, for example.

A phenotypic effect can be the causing or curing of a disease state, especially where mammalian cells are referred to herein. For cells of a pathogen or tumor cells, especially, a phenotypic effect can be the slowing of growth rate or cessation of growth.

Biomolecule: a molecule which can be produced as a gene product in cells that have been appropriately constructed to comprise one or more genes encoding the biomolecule. Preferably, production of the biomolecule can be turned on, when desired, by an inducible promoter. A biomolecule can be a peptide, polypeptide, or an RNA or RNA oligonucleotide, a DNA or DNA oligonucleotide, but is preferably a peptide. The same biomolecules can also be made synthetically. For peptides, see Merrifield, J., J. Am. Chem. Soc. 85: 2140-2154 (1963). For instance, an Applied Biosystems 431 A Peptide Synthesizer (Perkin Elmer) can be used for peptide synthesis. Biomolecules produced as gene products intracellularly are tested for their interaction with a target in the intracellular steps described herein (tests performed with cells in culture and tests performed with cells that have been introduced into animals). The same biomolecules produced synthetically are tested for their binding to an isolated target in an initial in vitro method described herein.

Synthetically produced biomolecules can also be used for a final step of the method for finding compounds that are competitive binders of the target.

Biomolecular Binder (of a target): a biomolecule which has been tested for its ability to bind to an isolated target cell component in vitro and has been found to bind to the target.

Biomolecular Inhibitor of Growth: a biomolecule which has been tested for its ability to inhibit the growth of cells constructed to produce the biomolecule in an "in culture" test of the effect of the biomolecule on growth of the cells, and has been found, in fact, to inhibit the growth of the cells in this test in culture.

Biomolecular Inhibitor of Infection: a biomolecule which has been tested for its ability to ameliorate the effects of infection, and has been found to do so. In the test, pathogen cells constructed to regulably express the biomolecule are introduced into one or more animals, the gene encoding the biomolecule is regulated so as to allow production of the biomolecule in the cells, and the effects of production of the biomolecule are observed in the infected animals compared to one or more suitable control animals.

Isolated: term used herein to indicate that the material in question exists in a physical milieu distinct from that in which it occurs in nature. For example, an isolated target cell component of the invention may be substantially isolated with respect to the complex cellular milieu in which it naturally occurs. The absolute level of purity is not critical, and those skilled in the art can readily determine appropriate levels of purity according to the use to which the material is to be put.

In many circumstances the isolated material will form part of a composition (for example, a more or less crude extract containing other substances), buffer system or reagent mix. In other circumstances, the material may be purified to essential homogeneity, for example as determined by PAGE or column chromatography (for example, HPLC).

Pathogen or Pathogenic Organism: an organism which is capable of causing disease, detectable by signs of infection or symptoms characteristic of disease. Pathogens can include procaryotes (which include, for example, medically significant Gram- positive bacteria such as Streptococcus pneumoniae, Enterococcus faecalis and Staphylococcus aureus, Gram-negative bacteria such as Escherichia coli, Pseudomonas aeroginosa and Klebsiella pneumoniae, and "acid-fast" bacteria such as Mycobacteria, especially M. tuberculosis), eucaryotes such as yeast and fungi (for example, Candida albicans and Aspergillus fumigatus) and parasites. It should be recognized that pathogens can include such organisms as soil-dwelling organisms and "normal flora" of the skin, gut and orifices, if such organisms colonize and cause symptoms of infection in a human or other mammal, by abnormal proliferation or by growth at a site from which the organism cannot usually be cultured.

Section 2. Whole cell engineering using real-time metabolic flux analysis

TECHNICAL FIELD

In one embodiment, the present invention provides methods for whole cell engineering, cell biology and molecular biology. In particular, the invention is directed to methods for whole cell engineering of new and modified phenotypes by using "on-line" or "real-time" metabolic flux analysis.

BACKGROUND

In one embodiment of this invention, whole cell metabolic flux analysis is a "horizontal" or "holistic" approach to study the metabolism, or "metabolome," of an organism. A whole cell "horizontal" metabolome approach studies the expression and function of all of the genes of an organism simultaneously. By using this whole cell approach to study a cell's metabolism, it is possible to get a complete snapshot of the whole cell's transcriptome (the expressed transcripts, or mRNA messages) and proteome (the expressed polypeptides). However, such snapshots are static pictures of one aspect of a cell's physiology and metabolism. Development of a means to dynamically monitor many different parameters in a cell culture would be much more effective in detecting new or altered cell phenotypes.

SUMMARY

One embodiment of this invention provides a method for whole cell engineering of new or modified phenotypes by using real-time metabolic flux analysis, the method comprising the following steps: (a) making a modified cell by modifying the genetic composition of a cell; (b) culturing the modified cell to generate a plurality of modified cells; (c) measuring at least one metabolic parameter of the cell by monitoring the cell culture of step (b) in real time; and, (d) analyzing the data of step (c) to determine if the measured parameter differs from a comparable measurement in an unmodified cell under similar conditions, thereby identifying an engineered phenotype in the cell using real-time metabolic flux analysis.

In one aspect, the genetic composition of the cell is modified by a method comprising addition of a nucleic acid to the cell. One or more nucleic acids can be added at the same time, or, in series. The genetic composition of the cell can be modified by addition of a nucleic acid heterologous to the cell, or, a nucleic acid homologous to the cell. The homologous nucleic acid can comprise a modified homologous nucleic acid, such as a modified homologous gene. The coding sequence or transcriptional regulatory sequence of a gene can be modified. Alternatively, the genetic composition of the cell can be modified by a method comprising deletion of a sequence or modification of a sequence in the cell. The genetic composition of the cell can be modified by a method comprising modifying or knocking out the expression of a gene.

The method can further comprising selecting a cell comprising a newly engineered phenotype. The selected cell can be isolated. The method can further comprise culturing the selected or isolated cell, thereby generating a new cell strain or cell line comprising a newly engineered phenotype. The methods can further comprise isolating a cell comprising a newly engineered phenotype.

Any phenotype can be added or modified. For example, a phenotype can be specifically targeted for change or addition. Thus, specific heterologous genes can be inserted or specific homologous genes can be stochastically or non-stochastically modified. For example, the newly engineered phenotype can be, e.g., an increased or decreased expression or amount of a polypeptide, an increased or decreased amount of an mRNA transcript, an increased or decreased expression of a gene, an increased or decreased resistance or sensitivity to a toxin, an increased or decreased resistance use or production of a metabolite, an increased or decreased uptake of a compound by the cell, an increased or decreased rate of metabolism, and an increased or decreased growth rate.

The newly engineered phenotype can a stable phenotype. In another aspect, it can be a transient or an inducible phenotype. In one aspect, modifying the genetic composition of a cell comprises insertion of a construct into the cell, wherein construct comprises a nucleic acid operably linked to a constitutively active promoter. Alternatively, modifying the genetic composition of a cell can comprise insertion of a construct into the cell, wherein construct comprises a nucleic acid operably linked to an inducible promoter. The nucleic acid added to the cell can be stably inserted into the genome of the cell. Alternatively, the nucleic acid added to the cell can propagate as an episome in the cell.

In one aspect, the nucleic acid added to the cell can encode a peptide or a polypeptide. The polypeptide can comprise a homologous polypeptide, such as a modified homologous polypeptide. Alternatively, the polypeptide can comprise a heterologous polypeptide. The nucleic acid added to the cell can encode a transcript comprising a sequence that is antisense to a homologous transcript. In one aspect, modifying the genetic composition of the cell can comprise increasing or decreasing the expression of an mRNA transcript. Modifying the genetic composition of the cell can comprise increasing or decreasing the expression of a polypeptide, a lipid, a mono- or poly-saccharide or a nucleic acid.

In one aspect, modifying the homologous gene can comprise knocking out expression of the homologous gene. Modifying the homologous gene can comprise increasing the expression of the homologous gene. The gene modification can be random, or stochastic, or, non-random, or targeted, i.e., non-stochastic.

In an exemplary non-stochastic gene modification, a gene to be inserted into a cell to modify a phenotype can be a heterologous gene or a sequence-modified homologous gene, wherein the sequence modification is made by a method comprising the following steps: (a) providing a template polynucleotide, wherein the template polynucleotide comprises a homologous gene of the cell (it can also be a heterologous gene that you wish to modify); (b) providing a plurality of oligonucleotides, wherein each oligonucleotide comprises a sequence homologous to the template polynucleotide, thereby targeting a specific sequence of the template polynucleotide, and a sequence that is a variant of the homologous gene; (c) generating progeny polynucleotides comprising non-stochastic sequence variations by replicating the template polynucleotide of step (a) with the oligonucleotides of step (b), thereby generating polynucleotides comprising homologous gene sequence variations. One variation of this method has been termed "gene site-saturation mutagenesis," "site-saturation mutagenesis," "saturation mutagenesis" or simply "GSSM," and is described in further detail, below. It can be used in combination with other mutagenization processes. See, e.g., U.S. Patent Nos. 6,171,820; 6,238,884.

Another exemplary non-stochastic gene modification process comprises introduction of two or more related polynucleotides into a suitable host cell such that a hybrid polynucleotide is generated by recombination and reductive reassortment. For example, the sequence modification of the gene to be modified (e.g., the heterologous gene or homologous gene) is made by a method comprising the following steps: (a) providing a template polynucleotide, wherein the template polynucleotide comprises sequence encoding a homologous gene; (b) providing a plurality of building block polynucleotides, wherein the building block polynucleotides are designed to cross-over reassemble with the template polynucleotide at a predetermined sequence, and a building block polynucleotide comprises a sequence that is a variant of the homologous gene and a sequence homologous to the template polynucleotide flanking the variant sequence; (c) combining a building block polynucleotide with a template polynucleotide such that the building block polynucleotide cross-over reassembles with the template polynucleotide to generate polynucleotides comprising homologous gene sequence variations. One variation of this method has been termed "synthetic ligation reassembly," or simply "SLR," and is described in further detail, below. It can be used in combination with other mutagenization processes. See, e.g., U.S. Patent No. 6,171,820.

Any cell can be engineered by the methods the invention, including, e.g., prokaryotic cells and eukaryotic cells. Bacteria, Archaebacteria, fungi, yeast, plant cells, insect cells, mammalian cells, including human cells, without limitation, can be engineered by the methods the invention. Furthermore, intracellular parasites, bacteria, viruses can be "indirectly" engineered by culturing and monitoring of eukaryotic cells by the methods the invention, including, e.g., immunodeficiency viruses, e.g., HIV, oncoviruses, mycobacteria, protozoan organisms (e.g., trypanosomes, such as *Trypanosoma rangeli*), plasmodium (e.g., *Plasmodium falciparum*), toxoplasmosis (e.g., *Toxoplasma gondii*), *Leishmania*, and the like.

In practicing the methods of the invention, any metabolic parameter can be measured. In one aspect, several different metabolic parameters are evaluated in the cell culture. The metabolic parameters can be measured at the same time or sequentially. One exemplary metabolic parameter is rate of cell growth, which can be measured by, e.g., a change in optical density of the cell culture. Another exemplary metabolic parameter measured comprises a change in the expression of a polypeptide. Changes in the expression of the polypeptide can be measured by any method, e.g., a one-dimensional gel electrophoresis, a two-dimensional gel electrophoresis, a tandem mass spectography, an RIA, an ELISA, an immunoprecipitation and a Western blot.

In one aspect, the measured metabolic parameter comprises a change in expression of at least one transcript, or, the expression of a transcript of a newly introduced gene. The change in expression of the transcript can be measured by a method selected from the group consisting of a hybridization, a quantitative amplification and a Northern blot. The transcript expression can be measured by hybridization of a sample comprising transcripts of a cell or nucleic acid representative of or complementary to transcripts of a cell by hybridization to immobilized nucleic acids on an array.

In one aspect, the measured metabolic parameter comprises a measurement of a metabolite, including primary and secondary metabolites. For example, the measured metabolic parameter can comprise an increase or a decrease in a primary or a secondary metabolite. The secondary metabolite can be selected from the group consisting of a glycerol and a methanol. The measured metabolic parameter can comprise an increase or a decrease in an organic acid, such as an acetate, a butyrate, a succinate and an oxaloacetate.

In one aspect, the measured metabolic parameter comprises an increase or a decrease in intracellular pH, or, extracellular pH in a culture medium. The increase or a decrease in intracellular pH can measured by intracellular application of a dye; the change in fluorescence of the dye can be measured over time. In one aspect, the measured metabolic parameter comprises gas exchange rate measurements.

In one aspect, the measured metabolic parameter comprises an increase or a decrease in synthesis of DNA or RNA over time. The increase or a decrease in synthesis, or accumulation, or decay, of DNA or RNA over time can be measured by intracellular application of a dye; the change in fluorescence of the dye can be measured over time.

In one aspect, the measured metabolic parameter comprises an increase or a decrease in uptake of a composition. The composition can be a metabolite, such as a monosaccharide, a disaccharide, a polysaccharide, a lipid, a nucleic acid, an amino acid and a polypeptide. The saccharide, disaccharide or polysaccharide can comprise a glucose or a sucrose. The composition can also be an antibiotic, a metal, a steroid and an antibody.

In one aspect, the measured metabolic parameter comprises an increase or a decrease in the secretion of a byproduct or a secreted composition of a cell. The byproduct or secreted composition can be a toxin, a lymphokine, a polysaccharide, a lipid, a nucleic acid, an amino acid, a polypeptide and an antibody.

In one aspect of the methods, the real time monitoring simultaneously measures a plurality of metabolic parameters. The real time monitoring of a plurality of metabolic parameters can comprise use of a Cell Growth Monitor device. The Cell Growth Monitor device can be a Wedgewood Technology, Inc., Cell Growth Monitor model 652, or similar model or variation thereof. In one aspect, the real time simultaneous monitoring measures uptake of substrates, levels of intracellular organic acids and levels of intracellular amino acids. The real time simultaneous monitoring can measure: uptake of glucose; levels of acetate, butyrate, succinate or oxaloacetate; and, levels of intracellular natural amino acids.

In one aspect, the method further comprises use of a computer-implemented program to real time monitor the change in measured metabolic parameters over time. The computer-implemented program can comprise a computer-implemented method as set forth in Figure 28. The computer-implemented method can comprise metabolic network equations. These computer-implemented method can also comprise a pathway analysis, an error analysis, such as a weighted least squares solution, and a flux estimation. The computer-implemented method can further comprises a preprocessing unit to filter out the errors for the measurement before the metabolic flux analysis.

The details of one or more aspects of the invention are set forth in the accompanying drawings and the description below. Other features, objects, and advantages of the invention will be apparent from the description and drawings, and from the claims.

All publications, GenBank Accession references (sequences), ATCC Deposits, patents and patent applications cited herein are hereby expressly incorporated by reference for all purposes.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 28 is a schematic illustrating an exemplary metabolic flux analysis (MFA) procedure of the invention.

DETAILED DESCRIPTION

In one embodiment, the invention provides novel methods for whole cell engineering of new and modified phenotypes by using “on-line” or “real-time” metabolic flux analysis. In practicing the methods of the invention, as a first step, a cell is modified by changing the genetic composition of the cell. The modification can be random, i.e., stochastic, or, by non-stochastic methods, as described herein. Specific genes or specific metabolic pathways can be targeted for modification.

In one aspect, the second step of the methods of the invention comprises culturing the modified cell to generate a plurality of modified cells. The cells can be cultured by any means, for example, in cell culture, such as a tissue culture, by fermentation or tissue culture reactors, or in a cell growth monitor device.

In one aspect, the next step of the methods comprises measuring at least one metabolic parameter of the cell in real time. In one aspect, a plurality of metabolome parameters are simultaneously measured. Thus, one or several devices can be used to monitor and measure metabolic parameters. For example, a cell growth monitor devices can measure a plurality of metabolic parameters of the cells in culture in real time. One example is the Wedgewood Technology, Inc. (San Carlos, CA), Cell Growth Monitor model 652™, as discussed below.

Finally, in one embodiment, the methods comprise analyzing these data of to determine if the measured parameters differ from a comparable measurement in an unmodified cell under similar conditions, or, change over time, thereby identifying an engineered phenotype in the cell using real-time metabolic flux analysis. For example, the parameter can be higher, lower or change at a rate that differs from a wild type cell or cell culture. It is not necessary to simultaneously monitor an unmodified cell or cell culture in real time to determine if and/or what phenotypic modifications result from the modification of the cell's genetic composition. Data and information already known can be used as a reference.

In one aspect of the invention, the methods further comprise use of a computer-implemented program to real time monitor the change in measured metabolic parameters over time and the analyze and display the resulting processed data. One exemplary computer-implemented program comprises a computer-implemented method as set forth in Figure 1. In this and other computer-implemented methods that can be used, the paradigm comprises use of metabolic network equations, metabolic pathway analyses, error analysis, such as a weighted least squares solution to give a flux estimation and the like.

In one aspect of the invention, a nucleic acid (or, the nucleic acid) responsible for the altered phenotype is identified, re-isolated, again modified (e.g., either stochastically or non-stochastically), reinserted into the cell, and the process of real-time metabolic flux analysis is iteratively repeated. The process can be iteratively repeated until a desired phenotype is engineered. For example, a plant cell and plant cell culture is subjected to iterative repetition of the methods of the invention until a new plant cell is made that comprises a desired new phenotype, e.g., enhanced growth, nutritional value or insect or drought resistance, or all or some of these characteristics. A pathogenic microorganism can be subjected to iterative repetition of the methods of the invention until it becomes non-pathogenic. A microorganism can be engineered to become lethal to another organism, such as an insect, or, to produce a variety of antibiotics or other compositions. Microorganisms can be subjected to iterative repetition of the methods of the invention to engineer, e.g., increased yield of desired products, removal of unwanted co-metabolites, improved utilization of inexpensive carbon and nitrogen sources, and adaptation to fermentor/ bioreactor growth conditions, increased production of a primary metabolite, increased production of a secondary metabolite, increased tolerance to acidic conditions, increased tolerance to basic conditions, increased tolerance to organic solvents, increased tolerance to high salt conditions and increased tolerance to high or low temperatures.

A complete biosynthetic pathway can be inserted into a cell. Any cell phenotype can be modified or any phenotype can be added to a cell using the methods of the invention, without limitation. The invention can be practiced in combination with other methods for inserting and screening for metabolic pathways, see, e.g., U.S. Patent No. 6,268,140, which describes producing and screening combinatorial metabolic libraries of multimeric proteins, or, U.S. Patent No. 5,712,146 , which describes vectors encoding polyketide synthases which in turn catalyze the production of a variety of polyketides.

DEFINITIONS

Unless defined otherwise, all technical and scientific terms used herein have the meaning commonly understood by a person skilled in the art to which this invention belongs. As used herein, the following terms have the meanings ascribed to them unless specified otherwise.

The terms "array" or "microarray" or "biochip" or "chip" as used herein is a plurality of target elements, each target element comprising a defined amount of one or more polypeptides or nucleic acids immobilized onto a defined area of a substrate surface, as discussed in further detail, below.

As used herein, the terms “computer” and “processor” are used in their broadest general contexts and incorporate all such devices, as described in detail, below.

The term “saturation mutagenesis” or “GSSM” includes a method that uses degenerate oligonucleotide primers to introduce point mutations into a polynucleotide, as described in detail, below.

The term “optimized directed evolution system” or “optimized directed evolution” includes a method for reassembling fragments of related nucleic acid sequences, e.g., related genes, and explained in detail, below.

The term “synthetic ligation reassembly” or “SLR” includes a method of ligating oligonucleotide fragments in a non-stochastic fashion, and explained in detail, below.

The term “antibody” includes a peptide or polypeptide derived from, modeled after or substantially encoded by an immunoglobulin gene or immunoglobulin genes, or fragments thereof, capable of specifically binding an antigen or epitope, see, e.g. Fundamental Immunology, Third Edition, W.E. Paul, ed., Raven Press, N.Y. (1993); Wilson (1994) J. Immunol. Methods 175:267-73; Yarmush (1992) J. Biochem. Biophys. Methods 25:85-97. The term antibody includes antigen-binding portions, i.e., “antigen binding sites,” (e.g., fragments, subsequences, complementarity determining regions (CDRs)) that retain capacity to bind antigen, including (i) a Fab fragment, a monovalent fragment consisting of the VL, VH, CL and CH1 domains; (ii) a F(ab')2 fragment, a bivalent fragment comprising two Fab fragments linked by a disulfide bridge at the hinge region; (iii) a Fd fragment consisting of the VH and CH1 domains; (iv) a Fv fragment consisting of the VL and VH domains of a single arm of an antibody, (v) a dAb fragment (Ward et al., (1989) Nature 341:544-546), which consists of a VH domain; and (vi) an isolated complementarity determining region (CDR). Single chain antibodies are also included by reference in the term "antibody."

Generating and Manipulating Nucleic Acids

The methods of the invention include modifying the genetic composition of a cell by addition of a heterologous nucleic acid into the cell or modification of a homologous gene in the cell. Nucleic acids can be isolated from a cell, recombinantly generated or made synthetically. The sequences can be isolated by, e.g., cloning and expression of cDNA libraries, amplification of message or genomic DNA by PCR, and the like. In practicing the methods of the invention, homologous genes can be modified by manipulating a template nucleic acid, as described herein. The invention can be practiced in conjunction with any method or protocol or device known in the art, which are well described in the scientific and patent literature.

General Techniques

The nucleic acids used to practice this invention, whether RNA, cDNA, genomic DNA, vectors, viruses or hybrids thereof, may be isolated from a variety of sources, genetically engineered, amplified, and/or expressed/ generated recombinantly. Recombinant polypeptides generated from these nucleic acids can be individually isolated or cloned and tested for a desired activity. Any recombinant expression system can be used, including bacterial, mammalian, yeast, insect or plant cell expression systems.

Alternatively, these nucleic acids can be synthesized *in vitro* by well-known chemical synthesis techniques, as described in, e.g., Adams (1983) J. Am. Chem. Soc. 105:661; Belousov (1997) Nucleic Acids Res. 25:3440-3444; Frenkel (1995) Free Radic. Biol. Med. 19:373-380; Blommers (1994) Biochemistry 33:7886-7896; Narang (1979) Meth. Enzymol. 68:90; Brown (1979) Meth. Enzymol. 68:109; Beaucage (1981) Tetra. Lett. 22:1859; U.S. Patent No. 4,458,066.

Techniques for the manipulation of nucleic acids, such as, e.g., subcloning, labeling probes (e.g., random-primer labeling using Klenow polymerase, nick translation, amplification), sequencing, hybridization and the like are well described in the scientific and patent literature, see, e.g., Sambrook, ed., MOLECULAR CLONING: A LABORATORY MANUAL (2ND ED.), Vols. 1-3, Cold Spring Harbor Laboratory, (1989); CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, Ausubel, ed. John Wiley & Sons, Inc., New York (1997); LABORATORY TECHNIQUES IN BIOCHEMISTRY AND MOLECULAR BIOLOGY: HYBRIDIZATION WITH NUCLEIC ACID PROBES, Part I. Theory and Nucleic Acid Preparation, Tijssen, ed. Elsevier, N.Y. (1993).

Nucleic acids, vectors, capsids, polypeptides, and the like can be analyzed and quantified by any of a number of general means well known to those of skill in the art. These include, e.g., analytical biochemical methods such as NMR, spectrophotometry, radiography, electrophoresis, capillary electrophoresis, high performance liquid chromatography (HPLC), thin layer chromatography (TLC), and hyperdiffusion chromatography, various immunological methods, e.g. fluid or gel precipitin reactions, immunodiffusion, immuno-electrophoresis, radioimmunoassays (RIAs), enzyme-linked immunosorbent assays (ELISAs), immuno-fluorescent assays, Southern analysis, Northern analysis, dot-blot analysis, gel electrophoresis (e.g., SDS-PAGE), nucleic acid or target or signal amplification methods, radiolabeling, scintillation counting, and affinity chromatography.

Another useful means of obtaining and manipulating nucleic acids used to practice the methods of the invention is to clone from genomic samples, and, if desired, screen and re-clone inserts isolated or amplified from, e.g., genomic clones or cDNA clones. Sources of nucleic acid used in the methods of the invention include genomic or cDNA libraries contained in, e.g., mammalian artificial chromosomes (MACs), see, e.g., U.S. Patent Nos. 5,721,118; 6,025,155; human artificial chromosomes, see, e.g., Rosenfeld (1997) Nat. Genet. 15:333-335; yeast artificial chromosomes (YAC); bacterial artificial chromosomes (BAC); P1 artificial chromosomes, see, e.g., Woon (1998) Genomics 50:306-316; P1-derived vectors (PACs), see, e.g., Kern (1997) Biotechniques 23:120-124; cosmids, recombinant viruses, phages or plasmids.

Amplification of Nucleic Acids

In practicing the methods of the invention, nucleic acids encoding heterologous or homologous, or modified nucleic acids, can be reproduced by, e.g., amplification. Amplification reactions can also be used to quantify the amount of nucleic acid in a sample (such as the amount of message in a cell sample), label the nucleic acid (e.g., to apply it to an array or a blot), detect the nucleic acid, or quantify the amount of a specific nucleic acid in a sample. In one aspect of the invention, message isolated from a cell or a cDNA library are amplified. The skilled artisan can select and design suitable oligonucleotide amplification primers. Amplification methods are also well known in the art, and include, *e.g.*, polymerase chain reaction, PCR (see, e.g., PCR PROTOCOLS, A GUIDE TO METHODS AND APPLICATIONS, ed. Innis, Academic Press, N.Y. (1990) and PCR STRATEGIES (1995), ed. Innis, Academic Press, Inc., N.Y., ligase chain reaction (LCR) (see, e.g., Wu (1989) Genomics 4:560; Landegren (1988) Science 241:1077; Barringer (1990) Gene 89:117); transcription amplification (see, e.g., Kwoh (1989) Proc. Natl. Acad.

Sci. USA 86:1173); and, self-sustained sequence replication (see, e.g., Guatelli (1990) Proc. Natl. Acad. Sci. USA 87:1874); Q Beta replicase amplification (see, e.g., Smith (1997) J. Clin. Microbiol. 35:1477-1491), automated Q-beta replicase amplification assay (see, e.g., Burg (1996) Mol. Cell. Probes 10:257-271) and other RNA polymerase mediated techniques (*e.g.*, NASBA, Cangene, Mississauga, Ontario); see also Berger (1987) Methods Enzymol. 152:307-316; Sambrook; Ausubel; U.S. Patent Nos. 4,683,195 and 4,683,202; Sooknanan (1995) Biotechnology 13:563-564.

Modification of Nucleic Acids

In practicing the methods of the invention, the genetic composition of a cell is altered by, e.g., modification of a homologous gene *ex vivo*, followed by its reinsertion into the cell. A homologous, heterologous or gene selected by the methods of the invention can be altered by any means, including, e.g., random or stochastic methods, or, non-stochastic, or "directed evolution," methods.

Methods for random mutation of genes are well known in the art, see, e.g., U.S. Patent No. 5,830,696. For example, mutagens can be used to randomly mutate a gene. Mutagens include, e.g., ultraviolet light or gamma irradiation, or a chemical mutagen, e.g., mitomycin, nitrous acid, photoactivated psoralens, alone or in combination, to induce DNA breaks amenable to repair by recombination. Other chemical mutagens include, for example, sodium bisulfite, nitrous acid, hydroxylamine, hydrazine or formic acid. Other mutagens are analogues of nucleotide precursors, e.g., nitrosoguanidine, 5-bromouracil, 2-aminopurine, or acridine. These agents can be added to a PCR reaction in place of the nucleotide precursor thereby mutating the sequence. Intercalating agents such as proflavine, acriflavine, quinacrine and the like can also be used.

Techniques in molecular biology can be used, e.g., random PCR mutagenesis, see, e.g., Rice (1992) Proc. Natl. Acad. Sci. USA 89:5467-5471; or, combinatorial multiple cassette mutagenesis, see, e.g., Crameri (1995) Biotechniques 18:194-196. Alternatively, nucleic acids, e.g., genes, can be reassembled after random, or "stochastic," fragmentation, see, e.g., U.S. Patent Nos. 6,291,242; 6,287,862; 6,287,861; 5,955,358; 5,830,721; 5,824,514; 5,811,238; 5,605,793.

Non-stochastic, or "directed evolution," methods include, e.g., saturation mutagenesis (GSSM), synthetic ligation reassembly (SLR), or a combination thereof. In one aspect of the invention, nucleic acids are selected, using real-time metabolic flux analysis, for conferring a new or modified phenotype on a cell, isolated, modified and reinserted into a cell to reiterate the steps of the methods of the invention. Polypeptides encoded by isolated and/or modified nucleic acids can be screened for an activity before their reinsertion into the cell by, e.g., using a capillary array platform. See, e.g., U.S. Patent Nos. 6,280,926; 5,939,250.

Saturation mutagenesis, or, GSSM

In one aspect of the invention, non-stochastic gene modification, a "directed evolution process," can be used to modify a gene to be inserted into a cell to add or modify a phenotype. Variations of this method have been termed "gene site-saturation mutagenesis," "site-saturation mutagenesis," "saturation mutagenesis" or simply "GSSM." It can be used in combination with other mutagenization processes. See, e.g., U.S. Patent Nos. 6,171,820; 6,238,884. In one aspect, GSSM comprises providing a template polynucleotide and a plurality of oligonucleotides, wherein each oligonucleotide comprises a sequence homologous to the template polynucleotide, thereby targeting a specific sequence of the template polynucleotide, and a sequence that is a variant of the homologous gene; generating progeny polynucleotides comprising non-stochastic sequence variations by replicating the template polynucleotide with the oligonucleotides, thereby generating polynucleotides comprising homologous gene sequence variations.

In one aspect, codon primers containing a degenerate N,N,G/T sequence are used to introduce point mutations into a polynucleotide, so as to generate a set of progeny polypeptides in which a full range of single amino acid substitutions is represented at each amino acid position, e.g., an amino acid residue in an enzyme active site or ligand binding site targeted to be modified. These oligonucleotides can comprise a contiguous first homologous sequence, a degenerate N,N,G/T sequence, and, optionally, a second homologous sequence. The downstream progeny translational products from the use of such oligonucleotides include all possible amino acid changes at each amino acid site along the polypeptide, because the degeneracy of the N,N,G/T sequence includes codons for all 20 amino acids.

In one aspect, one such degenerate oligonucleotide (comprised of, e.g., one degenerate N,N,G/T cassette) is used for subjecting each original codon in a parental polynucleotide template to a full range of codon substitutions. In another aspect, at least two degenerate cassettes are used – either in the same oligonucleotide or not, for subjecting at least two original codons in a parental polynucleotide template to a full range of codon substitutions. For example, more than one N,N,G/T sequence can be contained in one oligonucleotide to introduce amino acid mutations at more than one site. This plurality of N,N,G/T sequences can be directly contiguous, or separated by one or more additional nucleotide sequence(s). In another aspect, oligonucleotides serviceable for introducing additions and deletions can be used either alone or in combination with the codons containing an N,N,G/T sequence, to introduce any combination or permutation of amino acid additions, deletions, and/or substitutions.

In one aspect, simultaneous mutagenesis of two or more contiguous amino acid positions is done using an oligonucleotide that contains contiguous N,N,G/T triplets, i.e. a degenerate (N,N,G/T)n sequence. In another aspect, degenerate cassettes having less degeneracy than the N,N,G/T sequence are used. For example, it may be desirable in some instances to use (e.g. in an oligonucleotide) a degenerate triplet sequence comprised of only one N, where said N can be in the first second or third position of the triplet. Any other bases including any combinations and permutations thereof can be used in the remaining two positions of the triplet. Alternatively, it may be desirable in some instances to use (e.g. in an oligo) a degenerate N,N,N triplet sequence.

In one aspect, use of degenerate triplets (e.g., N,N,G/T triplets) allows for systematic and easy generation of a full range of possible natural amino acids (for a total of 20 amino acids) into each and every amino acid position in a polypeptide (in alternative aspects, the methods also include generation of less than all possible substitutions per amino acid residue, or codon, position). For example, for a 100 amino acid polypeptide, 2000 distinct species (i.e. 20 possible amino acids per position X 100 amino acid positions) can be generated. Through the use of an oligonucleotide or set of oligonucleotides containing a degenerate N,N,G/T triplet, 32 individual sequences can code for all 20 possible natural amino acids. Thus, in a reaction vessel in which a parental polynucleotide sequence is subjected to saturation mutagenesis using at least one such oligonucleotide, there are generated 32 distinct progeny polynucleotides encoding 20 distinct polypeptides. In contrast, the use of a non-degenerate oligonucleotide in site-directed mutagenesis leads to only one progeny polypeptide product per reaction vessel. Nondegenerate oligonucleotides can optionally be used in combination with degenerate primers disclosed; for example, nondegenerate oligonucleotides can be used to generate specific point mutations in a working polynucleotide. This provides one means to generate specific silent point mutations, point mutations leading to corresponding amino acid changes, and point mutations that cause the generation of stop codons and the corresponding expression of polypeptide fragments.

In one aspect, each saturation mutagenesis reaction vessel contains polynucleotides encoding at least 20 progeny polypeptide molecules such that all 20 natural amino acids are represented at the one specific amino acid position corresponding to the codon position mutagenized in the parental polynucleotide (other aspects use less than all 20 natural combinations). The 32-fold degenerate progeny polypeptides generated from each saturation mutagenesis reaction vessel can be subjected to clonal amplification (e.g. cloned into a suitable host, e.g., *E. coli* host, using, e.g., an expression vector) and subjected to expression screening. When an individual progeny polypeptide is identified by screening to display a favorable change in property (when compared to the parental polypeptide, such as increased affinity or avidity to an antigen), it can be sequenced to identify the correspondingly favorable amino acid substitution contained therein.

In one aspect, upon mutagenizing each and every amino acid position in a parental polypeptide using saturation mutagenesis as disclosed herein, favorable amino acid changes may be identified at more than one amino acid position. One or more new progeny molecules can be generated that contain a combination of all or part of these favorable amino acid substitutions. For example, if 2 specific favorable amino acid changes are identified in each of 3 amino acid positions in a polypeptide, the permutations include 3 possibilities at each position (no change from the original amino acid, and each of two favorable changes) and 3 positions. Thus, there are 3 x 3 x 3 or 27 total possibilities, including 7 that were previously examined - 6 single point mutations (i.e. 2 at each of three positions) and no change at any position.

In another aspect, site-saturation mutagenesis can be used together with another stochastic or non-stochastic means to vary sequence, e.g., synthetic ligation reassembly (see below), shuffling, chimerization, recombination and other mutagenizing processes and mutagenizing agents. This invention provides for the use of any mutagenizing process(es), including saturation mutagenesis, in an iterative manner.

Synthetic Ligation Reassembly (SLR)

Another non-stochastic gene modification, a "directed evolution process," that can be can be used in the methods of the invention to modify a gene to be inserted into a cell to add or modify a phenotype has been termed "synthetic ligation reassembly," or simply "SLR." SLR is a method of ligating oligonucleotide fragments together non-stochastically. This method differs from stochastic oligonucleotide shuffling in that the nucleic acid building blocks are not shuffled, concatenated or chimerized randomly, but rather are assembled non-stochastically. See, e.g., U.S. Patent Application Serial No. (USSN) 09/332,835 entitled "Synthetic Ligation Reassembly in Directed Evolution" and filed on June 14, 1999 ("USSN 09/332,835"). In one aspect, SLR comprises the following steps: (a) providing a template polynucleotide, wherein the template polynucleotide comprises sequence encoding a homologous gene; (b) providing a plurality of building block polynucleotides, wherein the building block polynucleotides are designed to cross-over reassemble with the template polynucleotide at a predetermined sequence, and a building block polynucleotide comprises a sequence that is a variant of the homologous gene and a sequence homologous to the template polynucleotide flanking the variant sequence; (c) combining a building block polynucleotide with a template polynucleotide such that the building block polynucleotide cross-over reassembles with the template polynucleotide to generate polynucleotides comprising homologous gene sequence variations.

SLR does not depend on the presence of high levels of homology between polynucleotides to be rearranged. Thus, this method can be used to non-stochastically generate libraries (or sets) of progeny molecules comprised of over 10^{100} different chimeras. SLR can be used to generate libraries comprised of over 10^{1000} different progeny chimeras. Thus, aspects of the present invention include non-stochastic methods of producing a set of finalized chimeric nucleic acid molecule shaving an overall assembly order that is chosen by design. This method includes the steps of generating by design a plurality of specific nucleic acid building blocks having serviceable mutually compatible ligatable ends, and assembling these nucleic acid building blocks, such that a designed overall assembly order is achieved.

The mutually compatible ligatable ends of the nucleic acid building blocks to be assembled are considered to be "serviceable" for this type of ordered assembly if they enable the building blocks to be coupled in predetermined orders. Thus the overall assembly order in which the nucleic acid building blocks can be coupled is specified by the design of the ligatable ends. If more than one assembly step is to be used, then the overall assembly order in which the nucleic acid building blocks can be coupled is also specified by the sequential order of the assembly step(s). In one aspect, the annealed building pieces are treated with an enzyme, such as a ligase (e.g. T4 DNA ligase), to achieve covalent bonding of the building pieces.

In one aspect, the design of the oligonucleotide building blocks is obtained by analyzing a set of progenitor nucleic acid sequence templates that serve as a basis for producing a progeny set of finalized chimeric polynucleotide molecules. These parental oligonucleotide templates thus serve as a source of sequence information that aids in the design of the nucleic acid building blocks that are to be mutagenized, e.g., chimerized or shuffled.

In one aspect of this method, the sequences of a plurality of parental nucleic acid templates are aligned in order to select one or more demarcation points. The demarcation points can be located at an area of homology, and are comprised of one or more nucleotides. These demarcation points are preferably shared by at least two of the progenitor templates. The demarcation points can thereby be used to delineate the boundaries of oligonucleotide building blocks to be generated in order to rearrange the parental polynucleotides. The demarcation points identified and selected in the progenitor molecules serve as potential chimerization points in the assembly of the final chimeric progeny molecules. A demarcation point can be an area of homology (comprised of at least one homologous nucleotide base) shared by at least two parental polynucleotide sequences. Alternatively, a demarcation point can be an area of homology that is shared by at least half of the parental polynucleotide sequences, or, it can be an area of homology that is shared by at least two thirds of the parental polynucleotide sequences. Even more preferably a serviceable demarcation points is an area of homology that is shared by at least three fourths of the parental polynucleotide sequences, or, it can be shared by at almost all of the parental polynucleotide sequences. In one aspect, a demarcation point is an area of homology that is shared by all of the parental polynucleotide sequences.

In one aspect, a ligation reassembly process is performed exhaustively in order to generate an exhaustive library of progeny chimeric polynucleotides. In other words, all possible ordered combinations of the nucleic acid building blocks are represented in the set of finalized chimeric nucleic acid molecules. At the same time, in another embodiment, the assembly order (i.e. the order of assembly of each building block in the 5' to 3 sequence of each finalized chimeric nucleic acid) in each combination is by design (or non-stochastic) as described above. Because of the non-stochastic nature of this invention, the possibility of unwanted side products is greatly reduced.

In another aspect, the ligation reassembly method is performed systematically. For example, the method is performed in order to generate a systematically compartmentalized library of progeny molecules, with compartments that can be screened systematically, e.g. one by one. In other words this invention provides that, through the selective and judicious use of specific nucleic acid building blocks, coupled with the selective and judicious use of sequentially stepped assembly reactions, a design can be achieved where specific sets of progeny products are made in each of several reaction vessels. This allows a systematic examination and screening procedure to be performed. Thus, these methods allow a potentially very large number of progeny molecules to be examined systematically in smaller groups.

Because of its ability to perform chimerizations in a manner that is highly flexible yet exhaustive and systematic as well, particularly when there is a low level of homology among the progenitor molecules, these methods provide for the generation of a library (or set) comprised of a large number of progeny molecules. Because of the non-stochastic nature of the instant ligation reassembly invention, the progeny molecules generated preferably comprise a library of finalized chimeric nucleic acid molecules having an overall assembly order that is chosen by design.

The saturation mutagenesis and optimized directed evolution methods also can be used to generate these amounts of different progeny molecular species.

It is appreciated that the invention provides freedom of choice and control regarding the selection of demarcation points, the size and number of the nucleic acid building blocks, and the size and design of the couplings. It is appreciated, furthermore, that the requirement for intermolecular homology is highly relaxed for the operability of this invention. In fact, demarcation points can even be chosen in areas of little or no intermolecular homology. For example, because of codon wobble, i.e. the degeneracy of codons, nucleotide substitutions can be introduced into nucleic acid building blocks without altering the amino acid originally encoded in the corresponding progenitor template. Alternatively, a codon can be altered such that the coding for an originally amino acid is altered. This invention provides that such substitutions can be introduced into the nucleic acid building block in order to increase the incidence of intermolecularly homologous demarcation points and thus to allow an increased number of couplings to be achieved among the building blocks, which in turn allows a greater number of progeny chimeric molecules to be generated.

In another aspect, the synthetic nature of the step in which the building blocks are generated allows the design and introduction of nucleotides (e.g., one or more nucleotides, which may be, for example, codons or introns or regulatory sequences) that can later be optionally removed in an *in vitro* process (e.g. by mutageneis) or in an *in vivo* process (e.g. by utilizing the gene splicing ability of a host organism). It is appreciated that in many instances the introduction of these nucleotides may also be desirable for many other reasons in addition to the potential benefit of creating a serviceable demarcation point.

Thus, according to another aspect, a nucleic acid building block can be used to introduce an intron. Thus, functional introns may be introduced into a man-made gene manufactured according to the methods described herein. The artificially introduced intron(s) can be functional in a host cells for gene splicing much in the way that naturally-occurring introns serve functionally in gene splicing.

Optimized Directed Evolution System

In practicing the methods of the invention, nucleic acids can also be modified by a method comprising an optimized directed evolution system. Optimized directed evolution is directed to the use of repeated cycles of reductive reassortment, recombination and selection that allow for the directed molecular evolution of nucleic acids through recombination. Optimized directed evolution allows generation of a large population of evolved chimeric sequences, wherein the generated population is significantly enriched for sequences that have a predetermined number of crossover events.

A crossover event is a point in a chimeric sequence where a shift in sequence occurs from one parental variant to another parental variant. Such a point is normally at the juncture of where oligonucleotides from two parents are ligated together to form a single sequence. This method allows calculation of the correct concentrations of oligonucleotide sequences so that the final chimeric population of sequences is enriched for the chosen number of crossover events. This provides more control over choosing chimeric variants having a predetermined number of crossover events.

In addition, this method provides a convenient means for exploring a tremendous amount of the possible protein variant space in comparison to other systems. Previously, if one generated, for example, 10^{13} chimeric molecules during a reaction, it would be extremely difficult to test such a high number of chimeric variants for a particular activity. Moreover, a significant portion of the progeny population would have a very high number of crossover events which resulted in proteins that were less likely to have increased levels of a particular activity. By using these methods, the population of chimerics molecules can be enriched for those variants that have a particular number of crossover events. Thus, although one can still generate 10^{13} chimeric molecules during a reaction, each of the molecules chosen for further analysis most likely has, for example, only three crossover events. Because the resulting progeny population can be skewed to have a predetermined number of crossover events, the boundaries on the functional variety between the chimeric molecules is reduced. This provides a more manageable number of variables when calculating which oligonucleotide from the original parental polynucleotides might be responsible for affecting a particular trait.

One method for creating a chimeric progeny polynucleotide sequence is to create oligonucleotides corresponding to fragments or portions of each parental sequence. Each oligonucleotide preferably includes a unique region of overlap so that mixing the oligonucleotides together results in a new variant that has each oligonucleotide fragment assembled in the correct order. Additional information can also be found in USSN 09/332,835. The number of oligonucleotides generated for each parental variant bears a relationship to the total number of resulting crossovers in the chimeric molecule that is ultimately created. For example, three parental nucleotide sequence variants might be provided to undergo a ligation reaction in order to find a chimeric variant having, for example, greater activity at high temperature. As one example, a set of 50 oligonucleotide sequences can be generated corresponding to each portions of each parental variant. Accordingly, during the ligation reassembly process there could be up to 50 crossover events within each of the chimeric sequences. The probability that each of the generated chimeric polynucleotides will contain oligonucleotides from each parental variant in alternating order is very low. If each oligonucleotide fragment is present in the ligation reaction in the same molar quantity it is likely that in some positions oligonucleotides from the same parental polynucleotide will ligate next to one another and thus not result in a crossover event. If the concentration of each oligonucleotide from each parent is kept constant during any ligation step in this example, there is a 1/3 chance (assuming 3 parents) that an oligonucleotide from the same parental variant will ligate within the chimeric sequence and produce no crossover.

Accordingly, a probability density function (PDF) can be determined to predict the population of crossover events that are likely to occur during each step in a ligation reaction given a set number of parental variants, a number of oligonucleotides corresponding to each variant, and the concentrations of each variant during each step in the ligation reaction. The statistics and mathematics behind determining the PDF is described below. By utilizing these methods, one can calculate such a probability density function, and thus enrich the chimeric progeny population for a predetermined number of crossover events resulting from a particular ligation reaction. Moreover, a target number of crossover events can be predetermined, and the system then programmed to calculate the starting quantities of each parental oligonucleotide during each step in the ligation reaction to result in a probability density function that centers on the predetermined number of crossover events.

These methods are directed to the use of repeated cycles of reductive reassortment, recombination and selection that allow for the directed molecular evolution of a nucleic acid encoding an polypeptide through recombination. This system allows generation of a large population of evolved chimeric sequences, wherein the generated population is significantly enriched for sequences that have a predetermined number of crossover events. A crossover event is a point in a chimeric sequence where a shift in sequence occurs from one parental variant to another parental variant. Such a point is normally at the juncture of where oligonucleotides from two parents are ligated together to form a single sequence. The method allows calculation of the correct concentrations of oligonucleotide sequences so that the final chimeric population of sequences is enriched for the chosen number of crossover events. This provides more control over choosing chimeric variants having a predetermined number of crossover events.

In addition, these methods provide a convenient means for exploring a tremendous amount of the possible protein variant space in comparison to other systems. By using the methods described herein, the population of chimerics molecules can be enriched for those variants that have a particular number of crossover events. Thus, although one can still generate 10^{13} chimeric molecules during a reaction, each of the molecules chosen for further analysis most likely has, for example, only three crossover events. Because the resulting progeny population can be skewed to have a predetermined number of crossover events, the boundaries on the functional variety between the chimeric molecules is reduced. This provides a more manageable number of variables when calculating which oligonucleotide from the original parental polynucleotides might be responsible for affecting a particular trait.

In one aspect, the method creates a chimeric progeny polynucleotide sequence by creating oligonucleotides corresponding to fragments or portions of each parental sequence. Each oligonucleotide preferably includes a unique region of overlap so that mixing the oligonucleotides together results in a new variant that has each oligonucleotide fragment assembled in the correct order. See also USSN 09/332,835.

The number of oligonucleotides generated for each parental variant bears a relationship to the total number of resulting crossovers in the chimeric molecule that is ultimately created. For example, three parental nucleotide sequence variants might be provided to undergo a ligation reaction in order to find a chimeric variant having, for example, greater activity at high temperature. As one example, a set of 50 oligonucleotide sequences can be generated corresponding to each portions of each parental variant. Accordingly, during the ligation reassembly process there could be up to 50 crossover events within each of the chimeric sequences. The probability that each of the generated chimeric polynucleotides will contain oligonucleotides from each parental variant in alternating order is very low. If each oligonucleotide fragment is present in the ligation reaction in the same molar quantity it is likely that in some positions oligonucleotides from the same parental polynucleotide will ligate next to one another and thus not result in a crossover event. If the concentration of each oligonucleotide from each parent is kept constant during any ligation step in this example, there is a 1/3 chance (assuming 3 parents) that a oligonucleotide from the same parental variant will ligate within the chimeric sequence and produce no crossover.

Accordingly, a probability density function (PDF) can be determined to predict the population of crossover events that are likely to occur during each step in a ligation reaction given a set number of parental variants, a number of oligonucleotides corresponding to each variant, and the concentrations of each variant during each step in the ligation reaction. The statistics and mathematics behind determining the PDF is described below. One can calculate such a probability density function, and thus enrich the chimeric progeny population for a predetermined number of crossover events resulting from a particular ligation reaction. Moreover, a target number of crossover events can be predetermined, and the system then programmed to calculate the starting quantities of each parental oligonucleotide during each step in the ligation reaction to result in a probability density function that centers on the predetermined number of crossover events.

Determining Crossover Events

Embodiments of the invention include a system and software that receive a desired crossover probability density function (PDF), the number of parent genes to be reassembled, and the number of fragments in the reassembly as inputs. The output of this program is a "fragment PDF" that can be used to determine a recipe for producing reassembled genes, and the estimated crossover PDF of those genes. The processing described herein is preferably performed in MATLAB® (The Mathworks, Natick, Massachusetts) a programming language and development environment for technical computing.

Iterative Processes

In practicing the methods of the invention, the process can be iteratively repeated. For example a nucleic acid (or, the nucleic acid) responsible for an altered phenotype is identified, re-isolated, again modified, reinserted into the cell, and the process of real-time metabolic flux analysis is iteratively repeated. The process can be iteratively repeated until a desired phenotype is engineered. For example, an entire biochemical pathway can be engineered into a cell. Any cell phenotype can be modified or any phenotype can be added to a cell using the methods of the invention, without limitation.

Nucleic acids can be modified using either stochastic or non-stochastic methods. In various aspects, the methods generate sets of chimeric nucleic acid and protein molecules, followed by insertion into a cell, culturing, and then screening by using real-time metabolic flux analysis for a particular activity, such as a changed or added desired phenotype. The invention is not limited to only a single round of screening. Based on this determination, a second round of reassembly can take place that enriches for progeny having a desired property or incurring a desired phenotype.

Similarly, if it is determined that a particular oligonucleotide has no affect at all on the desired trait (e.g., a new phenotype), it can be removed as a variable by synthesizing larger parental oligonucleotides that include the sequence to be removed. Since incorporating the sequence within a larger sequence prevents any crossover events, there will no longer be any variation of this sequence in the progeny polynucleotides. This iterative practice of determining which oligonucleotides are most related to the desired trait, and which are unrelated, allows more efficient exploration all of the possible protein variants that might be provide a particular trait or activity.

Automated Control of Reactions

The process of generating any of the reactions of the methods of the invention can be automated with the assistance of automated devices and robotic instruments. For example, in one aspect, a cell growth monitor device is used for real-time metabolic flux analysis, such as a Wedgewood Technology, Inc., Cell Growth Monitor model 652. As noted below, this device can be linked to a computer system. Another exemplary device is a TECAN GENESIS™ programmable robot made by Tecan Corporation (Hombrechtikon, Switzerland), which can be interfaced with a computer that determines the quantities of each oligonucleotide fragment to yield a resulting PDF. By linking a computer system that determines the proper quantities of each oligonucleotide to an automated robot, a complete ligation reassembly system is produced. Data links through serial or other interfaces will allow the data files generated from the ligation reassembly calculations to be forwarded in the proper format for the robotic system to automatically begin allocating the proper quantities of each oligonucleotide fragment into a reaction tube.

The automated system can include a plurality of oligonucleotide fragments derived from a series of nucleic acid sequence variants, wherein said fragments are configured to join one another at unique overhangs. The system also has a data input field configured to store a target number of crossover events in for each of the variant sequences.

Within the system is also a prediction module configured to determine the quantity of each of the fragments to admix together so that mixing the fragments results in a population of progeny molecules that are enriched for crossover events corresponding to the target number. The system also provides a robotic arm linked to the prediction module through a communication interface for automatically mixing the fragments in the determined quantities.

Mutagenized Oligonucleotides

While the optimized directed evolution method can use oligonucleotides that have a 100% fidelity to their parent polynucleotide sequence, this level of fidelity is not required. For example, if a set of three related parental polynucleotides are chosen to undergo ligation reassembly in order to create, e.g., a new phenotype, a set of oligonucleotides having unique overlapping regions can be synthesized by conventional methods. However a set of mutagenized oligonucleotides could also be synthesized. These mutagenized oligonucleotides are preferably designed to encode silent, conservative, or non-conservative amino acids.

The choice to enter a silent mutation might be made to, for example, add a region of nucleotide homology two fragments, but not affect the final translated protein. A non-conservative or conservative substitution is made to determine how such a change alters the function of the resultant polypeptide. This can be done if, for example, it is determined that mutations in one particular oligonucleotide fragment were responsible for increasing the activity of a peptide. By synthesizing mutagenized oligonucleotides (e.g.: those having a different nucleotide sequence than their parent), one can explore, in a controlled manner, how resulting modifications to the peptide or protein sequence affect the activity of the peptide or polypeptide.

Another method for creating variants of a nucleic acid sequence using mutagenized fragments includes first aligning a plurality of nucleic acid sequences to determine demarcation sites within the variants that are conserved in a majority of said variants, but not conserved in all of said variants. A set of first sequence fragments of the conserved nucleic acid sequences are then generated, wherein the fragments bind to one another at the demarcation sites. A second set of fragments of the not conserved nucleic acid sequences are then generated by, for example, a nucleic acid synthesizer. However, the not conserved, sequences are generated to have mutations at their demarcation site so that the second fragments have the same nucleotide sequence at the demarcation sites as said first fragments. This allows the not conserved sequences to still hybridize during the ligation reaction to the other parental sequences. Once the fragments are generated, a desired number of crossover events can be selected for each of the variants. The quantity of each of the first and second fragments is then calculated so that a ligation/incubation reaction between the calculated quantities of the first and second fragments will result in progeny molecules having the desired number of crossover events.

In Silico, or Computer, Models

In silico, or computer program-implemented, paradigms can be used in practicing the methods of the invention to design altered or new nucleic acids to modify cells for the creation of new phenotypes. One exemplary *in silico* method that can be used in practicing the methods of the invention for generating man-made polynucleotide sequences for the creation of new phenotypes detects shared domains between a plurality of template polynucleotides. It does so by aligning the template polynucleotides and identifying all sequence strings having a certain percentage of homology, e.g., about 75% to 95% sequence identity, that are shared between all of the template polynucleotides. This detects shared domains between the template polynucleotides. Next, domain sequences are switched from one template polynucleotide with the sequence of a corresponding domain. This is repeated until all domains have been switched with a corresponding domain on another template polynucleotide, thereby generating *in silico* a library of man-made polynucleotide sequences from a set of template polynucleotides.

In silico, or computer program-implemented, methods can also be used in practicing the methods of the invention to analyze metabolic flux data; see, e.g., Covert (2001) Trends Biochem. Sci. 26(3):179-186; Jamshidi (2001) Bioinformatics 17(3):286-287. For example, the quantitative relationship between a primary carbon source (e.g., for bacteria, acetate or succinate) uptake rate, oxygen uptake rate, and maximal cellular growth rate can be modeled *in silico*, and used complementary to the "real-time" or "on-line" monitoring of the invention, see, e.g., Edwards (2001) Nat. Biotechnol. 19(2):125-130. The effects of gene deletions in a central metabolic pathway can also be modeled *in silico*, and used complementary to the "real-time" or "on-line" monitoring of the invention, see, e.g., Edwards (2000) Proc. Natl. Acad. Sci. USA 97(10):5528-5533.

Measuring Metabolic Parameters

The methods of the invention involve whole cell evolution, or whole cell engineering, of a cell to develop a new cell strain having a new phenotype. To detect the new phenotype, at least one metabolic parameter of a modified cell is monitored in the cell in a "real time" or "on-line" time frame. In one aspect, a plurality of cells, such as a cell culture, is monitored in "real time" or "on-line." In one aspect, a plurality of metabolic parameters is monitored in "real time" or "on-line."

Metabolic flux analysis (MFA) is based on a known biochemistry framework. A linearly independent metabolic matrix is constructed based on the law of mass conservation and on the pseudo-steady state hypothesis (PSSH) on the intracellular metabolites. In practicing the methods of the invention, metabolic networks are established, including the:

·identity of all pathway substrates, products and intermediary metabolites

·identity of all the chemical reactions interconverting the pathway metabolites, the stoichiometry of the pathway reactions,

·identity of all the enzymes catalysing the reactions, the enzyme reaction kinetics,

·the regulatory interactions between pathway components, e.g. allosteric interactions, enzyme-enzyme interactions etc,

·intracellular compartmentalisation of enzymes or any other supramolecular organisation of the enzymes, and,

·the presence of any concentration gradients of metabolites, enzymes or effector molecules or diffusion barriers to their movement.

Once the metabolic network for a given strain is built, mathematic presentation by matrix notion can be introduced to estimate the intracellular metabolic fluxes if the on-line metabolome data is available.

Metabolic phenotype relies on the changes of the whole metabolic network within a cell. Metabolic phenotype relies on the change of pathway utilization with respect to environmental conditions, genetic regulation, developmental state and the genotype, etc. In one aspect of the methods of the invention, after the on-line MFA calculation, the dynamic behavior of the cells, their phenotype and other properties are analyzed by investigating the pathway utilization. For example, if the glucose supply is increased and the oxygen decreased during the yeast fermentation, the utilization of respiratory pathways will be reduced and/or stopped, and the utilization of the fermentative pathways will dominate. Control of physiological state of cell cultures will become possible after the pathway analysis. The methods of the invention can help determine how to manipulate the fermentation by determining how to change the substrate supply, temperature, use of inducers, etc. to control the physiological state of cells to move along desirable direction. In practicing the methods of the invention, the MFA results can also be compared with transcriptome and proteome data to design experiments and protocols for metabolic engineering or gene shuffling, etc.

In practicing the methods of the invention, any modified or new phenotype can be conferred and detected, including new or improved characteristics in the cell. Any aspect of metabolism or growth can be monitored.

Monitoring expression of an mRNA transcript

In one aspect of the invention, the engineered phenotype comprises increasing or decreasing the expression of an mRNA transcript or generating new transcripts in a cell. mRNA transcript, or message can be detected and quantified by any method known in the art, including, e.g., Northern blots, quantitative amplification reactions, hybridization to arrays, and the like. Quantitative amplification reactions include, e.g., quantitative PCR, including, e.g., quantitative reverse transcription polymerase chain reaction, or RT-PCR; quantitative real time RT-PCR, or "real-time kinetic RT-PCR" (see, e.g., Kreuzer (2001) Br. J. Haematol. 114:313-318; Xia (2001) Transplantation 72:907-914).

In one aspect of the invention, the engineered phenotype is generated by knocking out expression of a homologous gene. The gene's coding sequence or one or more transcriptional control elements can be knocked out, e.g., promoters enhancers. Thus, the expression of a transcript can be completely ablated or only decreased.

In one aspect of the invention, the engineered phenotype comprises increasing the expression of a homologous gene. This can be effected by knocking out of a negative control element, including a transcriptional regulatory element acting in *cis-* or *trans-* , or, mutagenizing a positive control element.

As discussed below in detail, one or more, or, all the transcripts of a cell can be measured by hybridization of a sample comprising transcripts of the cell, or, nucleic acids representative of or complementary to transcripts of a cell, by hybridization to immobilized nucleic acids on an array.

Monitoring expression of a polypeptides, peptides and amino acids

In one aspect of the invention, the engineered phenotype comprises increasing or decreasing the expression of a polypeptide or generating new polypeptides in a cell. Polypeptides, peptides and amino acids can be detected and quantified by any method known in the art, including, e.g., nuclear magnetic resonance (NMR), spectrophotometry, radiography (protein radiolabeling), electrophoresis, capillary electrophoresis, high performance liquid chromatography (HPLC), thin layer chromatography (TLC), hyperdiffusion chromatography, various immunological methods, e.g. immunoprecipitation, immunodiffusion, immuno-electrophoresis, radioimmunoassays (RIAs), enzyme-linked immunosorbent assays (ELISAs), immuno-fluorescent assays, gel electrophoresis (e.g., SDS-PAGE), staining with antibodies, fluorescent activated cell sorter (FACS), pyrolysis mass spectrometry, Fourier-Transform Infrared Spectrometry, Raman spectrometry, GC-MS, and LC-Electrospray and cap-LC-tandem-electrospray mass spectrometries, and the like. Novel bioactivities can also be screened using methods, or variations thereof, described in U.S. Patent No. 6,057,103. Furthermore, as discussed below in detail, one or more, or, all the polypeptides of a cell can be measured using a protein array.

Biosynthetically directed fractional ^{13}C labeling of proteinogenic amino acids can be monitored by feeding a mixture of uniformly ^{13}C-labeled and unlabeled carbon source compounds into a bioreaction network. Analysis of the resulting labeling pattern enables both a comprehensive characterization of the network topology and the determination of metabolic flux ratios of the amino acids; see, e.g., Szyperski (1999) Metab. Eng. 1:189-197.

Monitoring the expression of a metabolites and biosynthetic pathways

In one aspect, primary and secondary metabolites are the measured metabolic parameters. Any relevant primary and secondary metabolite can be monitored in real time. For example, the measured metabolic parameter can comprise an increase or a decrease in a primary or a secondary metabolite. The secondary metabolite can be, e.g., a glycerol or a methanol. The measured metabolic parameter can comprise an increase or a decrease in an organic acid, such as an acetate, a butyrate, a succinate and an oxaloacetate. In one aspect, the metabolic parameter measured comprises an increase or a decrease in an organic acid, such as an acetate, a butyrate, a succinate and an oxaloacetate.

The choice of which metabolite or metabolic or biosynthetic pathway to monitor "on-line" or in "real time" depends on which phenotype is desired to be added or modified. For example, limonene and other downstream metabolites of geranyl pyrophosphate can be monitored "on-line" or in "real time" as in U.S. Patent No. 6,291,745, which monitored to generate means for insect control in plants, see, e.g.,. Metabolites/ antibiotics in the supernatant in *Bacillus subtilis* can be monitored for effective insecticidal, antifungal and antibacterial agents, see, e.g., U.S. Patent No. 6,291,426. The methods of the invention can also be used to monitor metabolites of the tricarboxylic acid cycle and glycolysis, as in a *Bacillus subtilis* strain by Sauer (1997) Nat. Biotechnol. 15:448-452 (who also used fractional ^{13}C-labeling and two-dimensional nuclear magnetic resonance spectroscopy). The penicillin biosynthetic pathway can be monitored in real time in, e.g., *Penicillium chrysogenum*; see, e.g., Nielsen (1995) Biotechnol. Prog. 11(3):299-305; Jorgensen (1995) Appl. Microbiol. Biotechnol. 43(1):123-130. Asparagine linked (N-linked) glycosylation can be studied in real time; see, e.g., Nyberg (1999) Biotechnol. Bioeng. 62(3):336-347. The amount of amino acids liberated from peptides in cell cultures grown in a hydrolysate-supplemented medium can be studied in real time; see, e.g., Nyberg (1999) Biotechnol. Bioeng. 62(3):324-335, who studies pathway fluxes in Chinese hamster ovary cells grown in a complex (hydrolysate containing) medium. The methods of the invention can also be used to monitor flux distributions for maximal ATP production in mitochondria, including ATP yields for glucose, lactate, and palmitate; see, e.g., Ramakrishna (2001) Am. J. Physiol. Regul. Integr. Comp. Physiol. 280(3):R695-704. In bacteria, the methods of the invention can also be used to monitor seven essential reactions in the central metabolic pathways, glycolysis, pentose phosphate pathway, tricarboxylic acid cycle, for the growth in a glucose medium, e.g., glucose minimal media. For gene modification, the seven genes encoding these enzymes can be grouped into three categories: (1) pentose phosphate pathway genes, (2) three-carbon glycolytic genes, and (3) tricarboxylic acid cycle genes. See, e.g., Edwards (2000) Biotechnol. Prog. 16(6):927-939.

Monitoring intracellular pH

In one aspect, the increase or a decrease in intracellular pH is measured "on-line" or in "real time." The change in intracellular pH can be measured by intracellular application of a dye. The change in fluorescence of the dye can be measured over time.

Any system can be used to determine intracellular pH. If a dye if used, in one exemplary method, whole-field time-domain fluorescence lifetime imaging (FLIM) can be used. FLIM can be used for the quantitative imaging of concentration ratios of mixed fluorophores and quantitative imaging of perturbations to fluorophore environment; in FLIM, the image contrast is derived from the fluorescence lifetime at each point in a two-dimensional image (see, e.g., Cole (2001) J. Microsc. 203(Pt 3):246-257). Near-field scanning optical microscopy (NSOM) is a high-resolution scanning probe technique that can be used to obtain simultaneous optical and topographic images with spatial resolution of tens of nanometers (see, e.g., Kwak (2001) Anal. Chem. 73(14):3257-3262). A frequency domain fluorescence lifetime imaging microscope (FLIM) enables the measurement and reconstruction of three-dimensional nanosecond fluorescence lifetime images (see, e.g., Squire (1999) J. Microsc. 193(Pt 1):36-49).

Monitoring expression of gases

In one aspect, the measured metabolic parameter comprises gas exchange rate measurements. Any gas can be monitored, e.g., oxygen, carbon monoxide, carbon dioxide, nitrogen and the like. See, e.g., Follstad (1999) Biotechnol. Bioeng. 63(6):675-683.

Screening Methodologies and "On-line" Monitoring Devices

In practicing the methods of the invention, "real time" or "on-line" cell monitoring devices are used to identify an engineered phenotype in the cell using real-time metabolic flux analysis. Any screening method can be used in conjunction with these "real time" or "on-line" cell monitoring devices.

Cell growth monitor devices

In one aspect, real time monitoring of a plurality of metabolic parameters is done with use of a cell growth monitor device. One exemplary such device is a Wedgewood

Technology, Inc. (San Carlos, CA), Cell Growth Monitor model 652, which can "real time" or "on-line" monitor a variety of metabolic parameters, including: the uptake of substrates, such as glucose; the levels of intracellular intermediates, such as organic acids, e.g., acetate, butyrate, succinate, oxaloacetate; and, levels of amino acids. Any cell growth monitor device can be used, and these devices can be modified to measure any set of parameters, without limitation. Cell growth monitor device can be used in conjunction with any other measuring or monitoring devices, such as There are some rapid analysis of metabolites at the whole-cell level, using methods such as pyrolysis mass spectrometry, Fourier-Transform Infrared Spectrometry, Raman spectrometry, GC-MS, and LC-Electrospray and cap-LC-tandem-electrospray mass spectrometries.

Capillary Arrays

In addition to "biochip" arrays (see below), capillary arrays, such as the GIGAMATRIX™, Diversa Corporation, San Diego, CA, can be used to screen for or monitor a variety of compositions, including polypeptides, nucleic acids, metabolites, by-products, antibiotics, metals, and the like, without limitation. Capillary arrays provide another system for holding and screening samples. For example, a sample screening apparatus can include a plurality of capillaries formed into an array of adjacent capillaries, wherein each capillary comprises at least one wall defining a lumen for retaining a sample. The apparatus can further include interstitial material disposed between adjacent capillaries in the array, and one or more reference indicia formed within of the interstitial material. A capillary for screening a sample, wherein the capillary is adapted for being bound in an array of capillaries, can include a first wall defining a lumen for retaining the sample, and a second wall formed of a filtering material, for filtering excitation energy provided to the lumen to excite the sample.

A polypeptide or nucleic acid, e.g., a ligand, can be introduced into a first component into at least a portion of a capillary of a capillary array. Each capillary of the capillary array can comprise at least one wall defining a lumen for retaining the first component, and introducing an air bubble into the capillary behind the first component. A second component can be introduced into the capillary, wherein the second component is separated from the first component by the air bubble. A sample of interest can be introduced as a first liquid labeled with a detectable particle into a capillary of a capillary array, wherein each capillary of the capillary array comprises at least one wall defining a lumen for retaining the first liquid and the detectable particle, and wherein the at least one wall is coated with a binding material for binding the detectable particle to the at least one wall. The method can further include removing the first liquid from the capillary tube, wherein the bound detectable particle is maintained within the capillary, and introducing a second liquid into the capillary tube.

The capillary array can include a plurality of individual capillaries comprising at least one outer wall defining a lumen. The outer wall of the capillary can be one or more walls fused together. Similarly, the wall can define a lumen that is cylindrical, square, hexagonal or any other geometric shape so long as the walls form a lumen for retention of a liquid or sample. The capillaries of the capillary array can be held together in close proximity to form a planar structure. The capillaries can be bound together, by being fused (e.g., where the capillaries are made of glass), glued, bonded, or clamped side-by-side. The capillary array can be formed of any number of individual capillaries, for example, a range from 100 to 4,000,000 capillaries. A capillary array can form a microtiter plate having about 100,000 or more individual capillaries bound together.

Arrays, or "BioChips"

In one aspect of the invention, the monitored parameter is transcript expression. One or more, or, all the transcripts of a cell can be measured by hybridization of a sample comprising transcripts of the cell, or, nucleic acids representative of or complementary to transcripts of a cell, by hybridization to immobilized nucleic acids on an array, or "biochip." By using an "array" of nucleic acids on a microchip, some or all of the transcripts of a cell can be simultaneously quantified. Arrays comprising genomic nucleic acid can also be used to determine the genotype of a newly engineered strain made by the methods of the invention. "Polypeptide arrays" can also be used to simultaneously quantify a plurality of proteins.

The present invention can be practiced with any known "array," also referred to as a "microarray" or "nucleic acid array" or "polypeptide array" or "antibody array" or "biochip," or variation thereof. Arrays are generically a plurality of "spots" or "target elements," each target element comprising a defined amount of one or more biological molecules, e.g., oligonucleotides, immobilized onto a defined area of a substrate surface for specific binding to a sample molecule, e.g., mRNA transcripts.

In practicing the methods of the invention, known arrays and methods of making and using arrays can be incorporated in whole or in part, or variations thereof, as described, for example, in U.S. Patent Nos. 6,277,628; 6,277,489; 6,261,776; 6,258,606; 6,054,270; 6,048,695; 6,045,996; 6,022,963; 6,013,440; 5,965,452; 5,959,098; 5,856,174; 5,830,645; 5,770,456; 5,632,957; 5,556,752; 5,143,854; 5,807,522; 5,800,992; 5,744,305; 5,700,637; 5,556,752; 5,434,049; see also, e.g., WO 99/51773; WO 99/09217; WO 97/46313; WO 96/17958; see also, e.g., Johnston (1998) Curr. Biol. 8:R171-R174; Schummer (1997) Biotechniques 23:1087-1092; Kern (1997) Biotechniques 23:120-124; Solinas-Toldo (1997) Genes, Chromosomes & Cancer 20:399-407; Bowtell (1999) Nature Genetics Supp. 21:25-32. See also published U.S. patent applications Nos. 20010018642; 20010019827; 20010016322; 20010014449; 20010014448; 20010012537; 20010008765. The present invention can use any known array, e.g., GeneChips™, Affymetrix, Santa Clara, CA; SpectralChip™ Human BAC Arrays, Spectral Genomics, Houston, Texas; and their accompanying manufacturer's instructions.

Antibodies and Immunoblots

In practicing the methods of the invention, antibodies can be used to isolate, identify or quantify particular polypeptides or polysaccharides. The antibodies can be used in immunoprecipitation, staining (e.g., FACS), immunoaffinity columns, and the like. If desired, nucleic acid sequences encoding for specific antigens can be generated by immunization followed by isolation of polypeptide or nucleic acid, amplification or cloning and immobilization of polypeptide onto an array of the invention. Alternatively, the methods of the invention can be used to modify the structure of an antibody produced by a cell to be modified, e.g., an antibody's affinity can be increased or decreased. Furthermore, the ability to make or modify antibodies can be a phenotype engineered into a cell by the methods of the invention.

Methods of immunization, producing and isolating antibodies (polyclonal and monoclonal) are known to those of skill in the art and described in the scientific and patent literature, see, e.g., Coligan, CURRENT PROTOCOLS IN IMMUNOLOGY, Wiley/Greene, NY (1991); Stites (eds.) BASIC AND CLINICAL IMMUNOLOGY (7th ed.) Lange Medical

Publications, Los Altos, CA ("Stites"); Goding, MONOCLONAL ANTIBODIES: PRINCIPLES AND PRACTICE (2d ed.) Academic Press, New York, NY (1986); Kohler (1975) Nature 256:495; Harlow (1988) ANTIBODIES, A LABORATORY MANUAL, Cold Spring Harbor Publications, New York. Antibodies also can be generated *in vitro*, e.g., using recombinant antibody binding site expressing phage display libraries, in addition to the traditional *in vivo* methods using animals. See, e.g., Hoogenboom (1997) Trends Biotechnol. 15:62-70; Katz (1997) Annu. Rev. Biophys. Biomol. Struct. 26:27-45.

Sources of Cells and Culturing of Cells

The invention provides a method for whole cell engineering of new phenotypes by using real-time metabolic flux analysis. Any cell can be engineered, including, e.g., bacterial, Archaebacteria, mammalian, yeast, fungi, insect or plant cell. In one aspect of the methods of the invention, a cell is modified by addition of a heterologous nucleic acid into the cell. The heterologous nucleic acid can be isolated, cloned or reproduced from a nucleic acid from any source, including any bacterial, mammalian, yeast, insect or plant cell.

In one aspect, the cell can be from a tissue or fluid taken from an individual, e.g., a patient. The cell can be homologous, e.g., a human cell taken from a patient, or, heterologous, e.g., a bacterial or yeast cell taken from the gastrointestinal tract of an individual. The cell can be from, e.g., lymphatic or lymph node samples, serum, blood, chord blood, CSF or bone marrow aspirations, fecal samples, saliva, tears, tissue and surgical biopsies, needle or punch biopsies, and the like.

Any apparatus to grow or maintain cells can be used, e.g., a bioreactor or a fermentor, see, e.g., U.S. Patent Nos. 6,242,248; 6,228,607; 6,218,182; 6,174,720; 6,168,949; 6,133,022; 6,133,021; 6,048,721; 5,660,977; 5,075,234.

Real-time Metabolic Flux Analysis

In the methods of the invention, at least one metabolic parameter of the cell is monitored in real time, i.e., by real time, or "on-line," flux analysis. In alternative aspects, many parameters of the cells in culture are monitored simultaneously in real time. Because of the real-time distribution of substrates, intermediates and products between alternative metabolic pathways is not accessible by the usual analytical means, the present invention incorporates an MFA method with "on-line" or "real-time" metabolome data. Therefore, by calculation, the metabolic flux distributions during the fermentation can be quantified. The flux quantification and gene expression analysis, along with sophisticated experimental techniques, can be combined to upgrade the content of information in the physiological and genomic/proteomic data towards the unraveling of cellular function and regulation. This allows insight into metabolic pathways, which is highly desirable and necessary in order to understand the behavior of the organism.

Metabolic Flux Analysis (MFA) is an analysis technique for metabolic engineering. It has been used in connection with studies of cell metabolism where the aim is to direct as much carbon as possible from the substrate into the biomass and products. Example 1, below, generally describes an exemplary Metabolic Flux Analysis (MFA) that can be used in the methods of the invention..

"Metabolomics" is a relatively unexplored field and can encompass the analysis of all cellular metabolites. Metabolomics provides a powerful new tool for gaining insight into functional biology, and has provided snapshots of the levels of numerous small molecules within a cell, and how those levels change under different conditions. These studies are very complementary to gene and polypeptide expression studies (genomics and proteomics), which are actively being applied to studies of infectious diseases, production, and model organisms, as well as human cells and plants. The present invention provides an improved methodology to study "metabolomics" by providing a method for whole cell engineering of new or modified phenotypes by using real-time metabolic flux analysis.

In practicing the methods of the invention, cellular control can be studied at different hierarchical levels, at the level of the genome, at the level of the transcriptome, at the level of the proteome or at the level of the metabolome. Whilst there is much current interest in the genome-wide analysis of cells at the level of transcription (to define the 'transcriptome') and translation (to define the 'proteome'), the third level of analysis, that of the 'metabolome', has been curiously unexplored to date. The term 'metabolome' refers to the entire complement of all the small molecular weight metabolites inside a cell suspension (or other sample) of interest. It is likely that measurement of the metabolome in different physiological states, particularly using the methods of the invention, will in fact be much more discriminating for the purposes of functional genomics.

The genome (the total genetic material in the cell) specifies an organism's total repertoire of responses. The genomes of several organisms have now been completely sequenced and several others are near completion or well under way (including a number of parasites). Of the genes so far sequenced via the systematic genome sequencing programs, the functions of fewer than half are known with any confidence. Technological advances now allow gene expression at any particular stage of development or in any particular physiological state to be analyzed. Such analyses can be carried out at the level of transcription using either Northern blots or, more efficiently, using hybridization array technologies to determine which genes are being expressed under different sets of conditions, i.e., the "transcriptome." Similar analyses can be carried out at the level of translation to define the "proteome," i.e., the total protein complement of the cell. Improvements in 2D electrophoresis and computer software for advanced image analysis allow 1-2 x10^3 proteins to be resolved on a single 20x20 cm plate; and, mass spectrometry coupled with database searching provides a method for rapid protein identification. Changes in the transcriptome represent the initial response of a cell to change, while changes in the proteome represent the final response at the level of the macromolecule. The third level of analysis, and one analyzed by the methods of the invention, is that of the "metabolome," which includes the quantitative complement of all the low molecular weight molecules present in cells in a particular physiological or developmental state.

Metabolite levels, which are monitored in alternative aspects of the invention, are thus the variables of choice to measure in a quantitative analysis of cellular function. Metabolites represent the down stream amplification of changes occurring in the transcriptome or the proteome. Moreover, metabolites regulate gene expression through a network of feedback pathways such that metabolites drive expression and act as the link between the genome and metabolism. The number of metabolites in the metabolome is also lower, by about an order of magnitude than the number of gene products in the transcriptome or the proteome (a typical eukaryotic cell contains around 10^5 genes and 10^4 different expressed proteins but only about 10^3 different known metabolites). Therefore, in order to understand intermediary metabolism and to exploit this knowledge changes in the metabolome are much more relevant and will be much easier both to detect and to exploit than changes either in the transcriptome or the proteome.

The methods of the invention, by identifying sites of specific metabolic lesions via the metabolome, in addition to its inherent scientific interest, will lead to the detection of targets for potentially novel pharmaceuticals or agrochemicals in whole cells. The methods of the invention can also be used to design functional assays. From these results, they can enable the design of very much simpler assays in which only the targeted metabolites are studied for specific high throughput, mechanistic assays.

The metabolome analysis of the invention has the advantage of being an online non-invasive technology. While static metabolome analysis has some advantages over transcriptome and proteome analysis because, for many organisms, the number of metabolites was far fewer than the number of genes or proteins. However, static metabolome analysis had an intrinsic disadvantage as well. This was that while biochemistry could generate information about the metabolic pathways, there is no direct link between the metabolites and the genes. They were also problems in analysing the concentration or even the very presence of certain metabolites. Current identification technologies such as infra-red spectrometry, mass spectrometry, or nuclear magnetic resonance spectroscopy produced some information but their use was limited and could not properly analyze a living cell. The methods of the invention, by providing "online" or "real-time" non-invasive technology solved this problem. The "online" or "real-time" time dimension of the methods of the invention, lacking in older techniques is one important factor in the methods ability to analyze a living cell.

Metabolic flux analysis (MFA) is a powerful analysis tool that can couple observed extracellular phenomena, such as uptake/ excretion rates, growth rate, product and biomass yields, etc., with the intracellular carbon flux and energy distribution. The "on-line" or "real-time" MFA of the invention can be used to investigate the physiology of *Escherichia coli*, *Saccharomyces cerevisiae*, and hybridomas (see, e.g., Keasling (1998) Biotechnol. Bioeng. 5;58(2-3):231-239; Pramanik (1998) Biotechnol. Bioeng. 60(2):230-238; Nissen et al., 1997; Schulze et al., 1996; Follstad et al., 1999), lysine production and the effect of mutations in *Corynebacterium glutamicum* (see, e.g., Vallino (2000) Biotechnol. Bioeng. 67(6):872-885; Vallino and Stephanopoulos, 1993, 1994; Park et al., 1997; Dominguez (1998) Eur. J. Biochem. 254(1):96-102), riboflavin production in *Bacillus subtilis* (see, e.g., Sauer et al., 1996, 1998; Sauer (1997) Nat. Biotechnol. 15:448-452), penicillin production in

Penicillium chrysogenum (Nielsen (1995) Biotechnol. Prog. 11(3):299-305; Jorgensen (1995) Appl. Microbiol. Biotechnol. 43(1):123-130); and, peptide amino acid metabolism in Chinese hamster ovary (CHO) cells (see, e.g., Nyberg (1999) Biotechnol. Bioeng. 62(3):324-335; Nyberg (1999) Biotechnol. Bioeng. 62(3):336-347).

Moreover, the "on-line" or "real-time" MFA of the invention can be used in combination with NMR, MS, and/or GC–MS to yield hard to get information about futile cycles, the degree of reaction reversibility, as well as active pathways; see, e.g., Szyperski (1999) Metab. Eng. 1:189-197; Szyperski (1998) Q Rev. Biophys. 31:41-106; Szyperski (1995) Eur. J. Biochem. 232(2):433-448; Szyperski et al., 1997; Schmidt et al., 1998; Klapa (1999) Biotechnol. Bioeng. 62(4):375-391; Mollney et al., 1999; Park et al., 1999; Wiechert et al., 1999; Wittmann and Heinzle, 1999. Schilling, Edwards, and Palsson have even extended the use of MFA to include the analysis of genomic data and the structural properties of cellular networks (Schilling (2000-2001) Biotechnol. Bioeng. 71(4):286-306; Edwards and Palsson, 1998; Schilling et al., 1999a,b); to monitor the C(3)-C(4) metabolite interconversion at the anaplerotic node in many microorganisms (see, e.g., Petersen (2000) J. Biol. Chem. 275(46):35932-35941).

In MFA, the intracellular fluxes are calculated using a stoichiometric model for all the major intracellular reactions and by applying mass balances around the intracellular metabolites. As input to the calculations, a set of measured fluxes, typically the uptake rates of substrates and secretion rates of metabolites is used

The novel "real-time" or "on-line" metabolic flux analysis of the invention can provide data regarding a full suite of metabolites synthesized by a biological system under given environmental conditions and/or with genetic regulation. The "real-time" or "on-line" MFA methods of the invention can provide metabolomic data sets that are extremely complex. The MFA methods of the invention can be an adequate tool to handle, store, normalize, and evaluate the acquired data in order to describe the systemic response of a complex biological system. The Figure 1 is a schematic illustrating the invention's new application of MFA to determine new phenotypes, pathway utilizations and cell responses to the studied strains during actual cell culture or fermentation periods. The results can be either used for post-fermentation analysis, or immediate control of the metabolism.

The "on-line," or "real-time" methods of the invention can also incorporate other analytical devices, such as HPLC and GC/MS, to estimate flux distribution in metabolic networks (constructed with our biochemical knowledge and genomic/proteomic information database) from experimental measurements. With these devices, "snapshots" of the biological systems under study can be obtained periodically, e.g., about every 1, 5, 10, 15, 20, 25, or 30 minutes, depending on the number of metabolic parameters studied and number of devices used.

*Vector **r** for metabolome data*

The on-line MFA of the invention uses "rate of change" data, or the difference between current metabolic measurements and last measurements. The differences are calculated and stored in the "raw measurement" vector for error analysis before they can be used. Thus, in one aspect, a "preprocessing unit" is used to filter out the errors for the measurement before the metabolic flux analysis to make sure that quality data be used. See Example 1, below.

Computer Systems

In one aspect, the methods of the invention use computer-implemented methods/ programs to real time monitor the change in measured metabolic parameters over time. The methods of the invention can be practiced using any program language or computer / processor and in conjunction with any known software or methodology. For example, one of the programs called MATHEMATICA™ (Wolfram Research, Inc., Champaign, IL), such as MATHEMATICA 4.1™, or variations thereof, can be used, see Example 1, below; and, see also, e.g., Jamshidi (2001) Bioinformatics 17(3):286-287; Wilson (2001) Biophys. Chem. 91(3):281-304; Torrecilla (2001) J. Neurochem. 76(5):1291-1307.

The computer/ processor used to practice the methods of the invention can be a conventional general-purpose digital computer, e.g., a personal "workstation" computer, including conventional elements such as microprocessor and data transfer bus. The computer/ processor can further include any form of memory elements, such as dynamic random access memory, flash memory or the like, or mass storage such as magnetic disc optional storage.

For example, a conventional personal computer such as those based on an Intel microprocessor and running a Windows operating system can be used. Any hardware or software configuration can be used to practice the methods of the invention. For example, computers based on other well-known microprocessors and running operating system software such as UNIX, Linux, MacOS and others are contemplated.

EXAMPLES

The following examples are offered to illustrate, but not to limit the claimed invention.

Example 1: Metabolic Flux Analysis (MFA)

The following example describes implementation of an exemplary Metabolic Flux Analysis (MFA), which is applied in the real time analysis of cell cultures in the methods of the invention. Figure 1

Metabolic Flux Analysis (MFA) is important analysis technique of metabolic engineering. A flux balance can be written for each metabolite (y_i) within a metabolic system to yield the dynamic mass balance equations that interconnect the various metabolites. Generally, for a metabolic network that contains *m* compounds and *n* metabolic fluxes, all the transient material balances can be represented by a single matrix equation:

$dY/dt = A\,X(t) - r(t)$

where

Y: *m* dimensional vector of metabolite amounts per cell

X: *n* metabolic fluxes

A: Stoichiometric *m* x *n* matrix, and

r: vector of specific rates from measurements

The time constants characterizing metabolic transients are typically very rapid compared to the time constants of cell growth and process dynamics, therefore, the mass balances can be simplified to only consider the steady-state behavior. Eliminating the derivative yields: $A\,X(t) = r(t)$.

Provided that $m >= n$ and A is full rank, the weighted least squares solution of the above equation is: $X = (A^TA)^{-1}A^T r$.

The sensitivity of the solution can be investigated by the matrix: $dX/dr = (A^TA)^{-1}A^T$.

The elements of the above matrix are useful for the determination of the change of individual fluxes with respect to the error or perturbation in the measurements.

Inputs

Stoichiometric Equations

A stoichiometry matrix is derived from the chemical equations to be used in the analysis. The matrix consists of coefficients of chemical species involved in the reactions. Rows represent the species and columns represent the equations. For instance, if we consider the equations of energy production in cells:

$$2\ NADH + O2 + 6\ ADP \rightarrow 2\ NAD + 2\ H2O + 6\ ATP$$
$$2\ FADH + O2 + 4\ ADP \rightarrow 2\ FAD + 2\ H2O + 4\ ATP$$
$$ATP \rightarrow ADP$$

This system yields a stoichiometry matrix with 3 columns and as many rows as species to be considered in the overall system. In this case, 8 species are considered so the

NADH	-2	0	0
O2	-1	-1	0
NAD	2	0	0
H2O	2	2	0
FADH	0	-2	0
FAD	0	2	0
ATP	6	4	-1
ADP	-6	-4	1

matrix is 3 x 8.

Using these templates, the stoichiometric matrix is 35 x 33, and it is in the EXCEL 97™ file "stoichiex.xls". This is the matrix 'A' described above, and it is derived from the 33 chemical equations below.

1. CENTRAL METABOLIC PATHWAYS

1) GLC + ATP + NAD → 2 PYR + ADP + NADH + H2O
2) PYR + NADH → LAC + NAD
3) PYR + NAD → ACCOA + CO2 + NADH
4) ACCOA + OAA + NAD + H2O → AKG + CO2 + NADH
5) AKG + NAD → SUCCOA + CO2 + NADH
6) SUCCOA + ADP + H2O + FAD → FUM + ATP + FADH
7) FUM + H2O → MAL
8) MAL + NAD → OAA + NADH
9) GLN + ADP → GLU + NH3 + ATP
10) GLU + NAD → AKG + NH3 + NADH
11) MAL → PYR + CO2

2. BIOMASS SYNTHESIS: C50.5% H8.31% O32.93% N8.26%

12) 0.1016 GLC + 0.031 GLN + 0.008 ARG + 0.0003 ASN + 0.001 GLU + 0.0038 GLY + 0.0028 HIS + 0.0071 ILE +).008 LEU +).0043 LYS + 0.001 MET + 0.0152 THR +).0051 VAL → BIOMASS

3. AMINO ACID METABOLISM

13) PYR + GLU → ALA + AKG
14) SER → PYR + NH3
15) GLY → SER
16) CYS → PYR + NH3
17) ASP + AKG → OAA + GLU
18) ASN → ASP + NH3
19) HIS → GLU + NH3
20) ARG + AKG → 2 GLU
21) PRO → GLU
22) ILE + AKG → SUCCOA + ACCOA + GLU
23) VAL + AKG → GLU + CO2 + SUCCOA
24) MET → SUCCOA
25) THR → SUCCOA + NH3
26) PHE → TYR
27) TYR + AKG → GLU + FUM + 2 ACCOA
28) LYS + 2 AKG → 2 GLU + 2 CO2 + 2 ACCOA
29) LEU + AKG → GLU + 3 ACCOA

4. ANTIBODY FORMATION:

30) 1.05 ARG +1.98 ASN + 1.96 ASP + 1.42 GLU +1.31 GLY +1.59 ILE + 3.79 LEU + 1.97 LYS +0.67 MET + 0.95 PHE + 5.72 SER 1.32 THR 5.05 TYR +2.68 VAL → Ab

5. ENERGY PRODUCTION:

31) 2 NADH + O2 + 6 ADP → 2 NAD + 2 H2O + 6 ATP
32) 2 FADH + O2 + 4 ADP → 2 FAD + 2 H2O + 4 ATP
33) ATP → ADP

In order to use this matrix with other mathematics software, it must be converted to a text file. Highlight only the cells that contain numbers, select copy from the Edit menu, and paste into a notepad (or simple text editor) document, e.g., the "Notepad" text editor program that comes with Microsoft Windows™ 3.11, 95 and NT. The file can be saved in a notepad as a text file "*.txt".

Specific Uptake Rates

The specific uptake rates are calculated from data from a cell culture reactor. This data should also be in a text file as a vector of rates, **r**, that correspond to the appropriate chemical species, i.e. the rows in the stoichiometry matrix above. In the provided templates, the specific rates are listed in the EXCEL 97™ file "ratex.xls" as well as a text file (exported from Excel) "rate.txt".

Calculations

With the inputs in the desired form, it is now time to use a mathematics software package to calculate the estimated internal fluxes. This software should be able to handle matrix math and differential equations. One template was made in MATHEMATICA™ 3.0 and is named "mfamath.nb". The following section assumes that the calculations are done in MATHEMATICA™ 3.0, but the general procedure can be applied with any suitable package.

Read in Data

First the default directory is set using the SetDirectory command:

```
example: SetDirectory["a:\mfa\"]
```

The data is then read in and saved into the A matrix (for the stoichiometry matrix) and the r vector (for the specific rates).

```
example: A=ReadList["stoichi.txt, Number, RecordLists --> True]
         r = ReadList["rate.txt, Number, RecordLists --> True]
```

Sensitivity Analysis

Next, the sensitivity matrix (dX/dr) is calculated as $(A^TA)^{-1}A^T$.

```
example: sens = Inverse[Transpose[A].A].Transpose[A]
```

Solution and Error Analysis

The least squares estimation of the flux distributions, x, and the errors, e, are calculated for the over-determined system of equations.

```
example: x = sens.r
         e = r - A.x
```

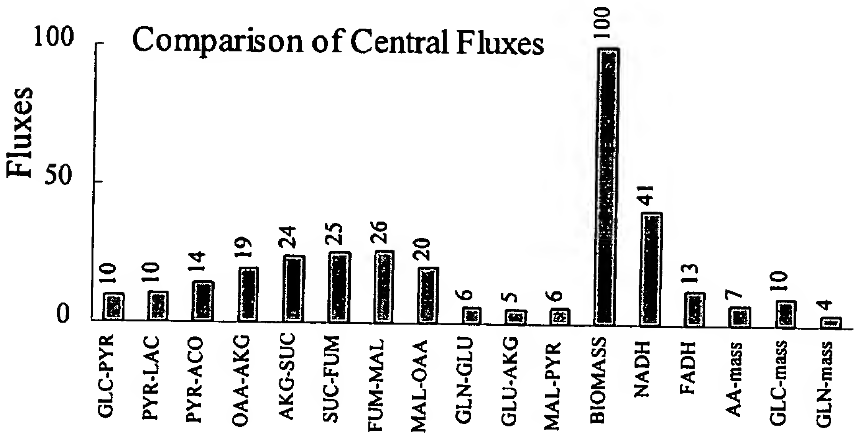
Output of Results

After calculation of the flux estimations, the results must be written to text files for presentation. In the templates provided, 3 results text files are included. These files are “flux.txt” that contains the x vector, “error.txt” that holds the error vector, and “sensitivity.txt” that contains the sensitivity matrix. An example of creating these text files in MATHEMATICA™ is shown below.

```
    Example: a1 = OpenWrite["flux.txt". FormatType ->
OutputForm];
    Write[a1, TableForm[x, TableSpacing -> {0,1}]]; Close[a1]
```

Presentation of Results

A critical aspect of this analysis is the efficient and clear presentation of the large number of estimated fluxes. The output text files from MATHEMATICA™ can be imported into Excel, and the solution can be plotted as a collection of bar graphs.



The EXCEL 97™ file "mfaexc.xls" is the template provided that shows the table of data and the bar graphs for each flux. It also contains a composite bar graph that plots the fluxes together and grouped by metabolic pathway (see below).

An additional way to present the data is to show all the internal fluxes overlain on a map of the relevant metabolic pathways. The POWERPOINT™ template file "mfa.ppt" shows a metabolic map with bar graphs (linked to the Excel file "mfaexc.xls" which must be opened before the file "mfa.ppt") to show the magnitude of the fluxes. There exists a linking between the Excel file and the POWERPOINT™ presentation. When the data in Excel is updated, the linking in the presentation should be updated.

Devices to monitor organic acids and amino acids

On-line devices that can monitor organic acids and amino acids can also be used in practicing the methods of the invention. For example, in one aspect, the BIO+ ON-LINE™ (Lachat Instruments, Milwaukee, WI) provides near-real-time monitoring of fermentation and mammalian cell culture processes. This device can provide critical information to maximize product yields. Mounted on a cart, this device can be rolled up to a fermentation bank and connected via a stream selector valve. From there, chemical constituent monitoring occurs automatically for ammonia, glucose, glutamate, glutamine, glycerol, lactate and phosphate individually and organic acids as a profile employing ion exclusion chromatography. The BIO+ ON-LINE™ is an integrated sampling system that provides a real solution to this challenging problem using a pumping system combined with a FLOWNAMICS® filter probe which exhibits the following benefits: sterilizable in-place; risk-free sampling due to elimination of bypass filters which recirculate material back into the vessel; sterile, cell-free sampling; accommodates all vessel sizes; minimum dead volume to ensure consistent and accurate sampling and to reduce flush time; durable design and construction to withstand temperatures, pressures, viscosities, shear forces and chemical constituents typical of bioprocess environments.

The BIO+ ON-LINE™ can determine up to four analytes simultaneously using flow injection analysis. The reaction modules can be removed and substituted with other modules. Thus, the user can customize the unit for different fermentation/ bioprocess requirements. Additionally, the Ion Chromatography channel can be customized to meet other Liquid Chromatography (LC) needs. While conductivity detection is the default detector, users can connect UV, RI, or other detectors and their own columns to the unit to meet their customized LC separation needs. This system, or variations thereof, is applicable to aerobic and anaerobic bacterial cultures as well as yeast, fungi, algae, insect and mammalian cell cultures.

Other related devices that can be used to practice the invention include the QUIKCHEM® 8000 (Lachat Instruments, Milwaukee, WI) which allows high sample throughput coupled with simple and rapid method changeover to maximize productivity in determining ionic species in a diversity of sample matrices from sub-ppb to percent concentrations.

One skilled in the art will readily appreciate that the present invention is well adapted to carry out the objects and obtain the ends and advantages mentioned as well as those inherent therein. The methods described herein are presently representative of exemplary aspects and are not intended as limitations on the scope of the invention. Changes therein and other uses will occur to those skilled in the art which are encompassed within the spirit of the invention and are defined by the scope of the claims.

3. MODIFYING : DIRECTED EVOLUTION METHODS

In one aspect the invention described herein is directed to the use of repeated cycles of reductive reassortment, recombination and selection which allow for the directed molecular evolution of highly complex linear sequences, such as DNA, RNA or proteins thorough recombination.

In vivo shuffling of molecules can be performed utilizing the natural property of cells to recombine multimers. While recombination *in vivo* has provided the major natural route to molecular diversity, genetic recombination remains a relatively complex process that involves 1) the recognition of homologies; 2) strand cleavage, strand invasion, and metabolic steps leading to the production of recombinant chiasma; and finally 3) the resolution of chiasma into discrete recombined molecules. The formation of the chiasma requires the recognition of homologous sequences.

In a preferred embodiment, the invention relates to a method for producing a hybrid polynucleotide from at least a first polynucleotide and a second polynucleotide. The present invention can be used to produce a hybrid polynucleotide by introducing at least a first polynucleotide and a second polynucleotide which share at least one region of partial sequence homology into a suitable host cell. The regions of partial sequence homology promote processes which result in sequence reorganization producing a hybrid polynucleotide. The term "hybrid polynucleotide", as used herein, is any nucleotide sequence which results from the method of the present invention and contains sequence from at least two original polynucleotide sequences. Such hybrid polynucleotides can result from intermolecular recombination events which promote sequence integration between DNA molecules. In addition, such hybrid polynucleotides can result from intramolecular reductive reassortment processes which utilize repeated sequences to alter a nucleotide sequence within a DNA molecule.

The invention provides a means for generating hybrid polynucleotides which may encode biologically active hybrid polypeptides. In one aspect, the original polynucleotides encode biologically active polypeptides. The method of the invention produces new hybrid polypeptides by utilizing cellular processes which integrate the sequence of the original polynucleotides such that the resulting hybrid polynucleotide encodes a polypeptide demonstrating activities derived from the original biologically active polypeptides. For example, the original polynucleotides may encode a particular enzyme from different microorganisms. An enzyme encoded by a first polynucleotide from one organism may, for example, function effectively under a particular environmental condition, e.g. high salinity. An enzyme encoded by a second polynucleotide from a different organism may function effectively under a different environmental condition, such as extremely high temperatures. A hybrid polynucleotide containing sequences from the first and second original polynucleotides may encode an enzyme which exhibits characteristics of both enzymes encoded by the original polynucleotides. Thus, the enzyme encoded by the hybrid polynucleotide may function effectively under environmental conditions shared by each of the enzymes encoded by the first and second polynucleotides, e.g., high salinity and extreme temperatures.

Enzymes encoded by the original polynucleotides of the invention include, but are not limited to; oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. A hybrid polypeptide resulting from the method of the invention may exhibit specialized enzyme activity not displayed in the original enzymes. For example, following recombination and/or reductive reassortment of polynucleotides encoding hydrolase activities, the resulting hybrid polypeptide encoded by a hybrid polynucleotide can be screened for specialized hydrolase activities obtained from each of the original enzymes, i.e. the type of bond on which the hydrolase acts and the temperature at which the hydrolase functions. Thus, for example, the hydrolase may be screened to ascertain those chemical functionalities which distinguish the hybrid hydrolase from the original hydrolyases, such as: (a) amide (peptide bonds), i.e. proteases; (b) ester bonds, i.e. esterases and lipases; (c) acetals, i.e., glycosidases and, for example, the temperature, pH or salt concentration at which the hybrid polypeptide functions.

Sources of the original polynucleotides may be isolated from individual organisms ("isolates"), collections of organisms that have been grown in defined media ("enrichment cultures"), or, most preferably, uncultivated organisms ("environmental samples"). The use of a culture-independent approach to derive polynucleotides encoding novel bioactivities from environmental samples is most preferable since it allows one to access untapped resources of biodiversity.

"Environmental libraries" are generated from environmental samples and represent the collective genomes of naturally occurring organisms archived in cloning vectors that can be propagated in suitable prokaryotic hosts. Because the cloned DNA is initially extracted directly from environmental samples, the libraries are not limited to the small fraction of prokaryotes that can be grown in pure culture. Additionally, a normalization of the environmental DNA present in these samples could allow more equal representation of the DNA from all of the species present in the original sample. This can dramatically increase the efficiency of finding interesting genes from minor constituents of the sample which may be under-represented by several orders of magnitude compared to the dominant species.

For example, gene libraries generated from one or more uncultivated microorganisms are screened for an activity of interest. Potential pathways encoding bioactive molecules of interest are first captured in prokaryotic cells in the form of gene expression libraries. Polynucleotides encoding activities of interest are isolated from such libraries and introduced into a host cell. The host cell is grown under conditions which promote recombination and/or reductive reassortment creating potentially active biomolecules with novel or enhanced activities.

The microorganisms from which the polynucleotide may be prepared include prokaryotic microorganisms, such as Eubacteria and Archaebacteria, and lower eukaryotic microorganisms such as fungi, some algae and protozoa. Polynucleotides may be isolated from environmental samples in which case the nucleic acid may be recovered without culturing of an organism or recovered from one or more cultured organisms. In one aspect, such microorganisms may be extremophiles, such as hyperthermophiles, psychrophiles, psychrotrophs, halophiles, barophiles and acidophiles. Polynucleotides encoding enzymes isolated from extremophilic microorganisms are particularly preferred. Such enzymes may function at temperatures above 100°C in terrestrial hot springs and deep sea thermal vents, at temperatures below 0°C in arctic waters, in the saturated salt environment of the Dead Sea, at pH values around 0 in coal deposits and geothermal sulfur-rich springs, or at pH values greater than 11 in sewage sludge. For example, several esterases and lipases cloned and expressed from extremophilic organisms show high activity throughout a wide range of temperatures and pHs.

Polynucleotides selected and isolated as hereinabove described are introduced into a suitable host cell. A suitable host cell is any cell which is capable of promoting recombination and/or reductive reassortment. The selected polynucleotides are preferably already in a vector which includes appropriate control sequences. The host cell can be a higher eukaryotic cell, such as a mammalian cell, or a lower eukaryotic cell, such as a yeast cell, or preferably, the host cell can be a prokaryotic cell, such as a bacterial cell. Introduction of the construct into the host cell can be effected by calcium phosphate transfection, DEAE-Dextran mediated transfection, or electroporation (**Davis** et al, 1986).

As representative examples of appropriate hosts, there may be mentioned: bacterial cells, such as *E. coli, Streptomyces, Salmonella typhimurium*; fungal cells, such as yeast; insect cells such as *Drosophila S2* and *Spodoptera Sf9*; animal cells such as CHO, COS or Bowes melanoma; adenoviruses; and plant cells. The selection of an appropriate host is deemed to be within the scope of those skilled in the art from the teachings herein.

With particular references to various mammalian cell culture systems that can be employed to express recombinant protein, examples of mammalian expression systems include the COS-7 lines of monkey kidney fibroblasts, described in "SV40-transformed simian cells support the replication of early SV40 mutants" (**Gluzman,** 1981), and other cell lines capable of expressing a compatible vector, for example, the C127, 3T3, CHO, HeLa and BHK cell lines. Mammalian expression vectors will comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences. DNA sequences derived from the SV40 splice, and polyadenylation sites may be used to provide the required nontranscribed genetic elements.

Host cells containing the polynucleotides of interest can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants or amplifying genes. The culture conditions, such as temperature, pH and the like, are those previously used with the host cell selected for expression, and will be apparent to the ordinarily skilled artisan. The clones which are identified as having the specified enzyme activity may then be sequenced to identify the polynucleotide sequence encoding an enzyme having the enhanced activity.

In another aspect, it is envisioned the method of the present invention can be used to generate novel polynucleotides encoding biochemical pathways from one or more operons or gene clusters or portions thereof. For example, bacteria and many eukaryotes have a coordinated mechanism for regulating genes whose products are involved in related processes. The genes are clustered, in structures referred to as "gene clusters," on a single chromosome and are transcribed together under the control of a single regulatory sequence, including a single promoter which initiates transcription of the entire cluster. Thus, a gene cluster is a group of adjacent genes that are either identical or related, usually as to their function. An example of a biochemical pathway encoded by gene clusters are polyketides. Polyketides are molecules which are an extremely rich source of bioactivities, including antibiotics (such as tetracyclines and erythromycin), anti-cancer agents (daunomycin), immunosuppressants (FK506 and rapamycin), and veterinary products (monensin). Many polyketides (produced by polyketide synthases) are valuable as therapeutic agents. Polyketide synthases are multifunctional enzymes that catalyze the biosynthesis of an enormous variety of carbon chains differing in length and patterns of functionality and cyclization. Polyketide synthase genes fall into gene clusters and at least one type (designated type I) of polyketide synthases have large size genes and enzymes, complicating genetic manipulation and *in vitro* studies of these genes/proteins.

The ability to select and combine desired components from a library of polyketides, or fragments thereof, and postpolyketide biosynthesis genes for generation of novel polyketides for study is appealing. The method of the present invention makes it possible to facilitate the production of novel polyketide synthases through intermolecular recombination.

Preferably, gene cluster DNA can be isolated from different organisms and ligated into vectors, particularly vectors containing expression regulatory sequences which can control and regulate the production of a detectable protein or protein-related array activity from the ligated gene clusters. Use of vectors which have an exceptionally large capacity for exogenous DNA introduction are particularly appropriate for use with such gene clusters and are described by way of example herein to include the f-factor (or fertility factor) of *E. coli*. This f-factor of *E. coli* is a plasmid which affect high-frequency transfer of itself during conjugation and is ideal to achieve and stably propagate large DNA fragments, such as gene clusters from mixed microbial samples. Once ligated into an appropriate vector, two or more vectors containing different polyketide synthase gene clusters can be introduced into a suitable host cell. Regions of partial sequence homology shared by the gene clusters will promote processes which result in sequence reorganization resulting in a hybrid gene cluster. The novel hybrid gene cluster can then be screened for enhanced activities not found in the original gene clusters.

Therefore, in a preferred embodiment, the present invention relates to a method for producing a biologically active hybrid polypeptide and screening such a polypeptide for enhanced activity by:

1) introducing at least a first polynucleotide in operable linkage and a second polynucleotide in operable linkage, said at least first polynucleotide and second polynucleotide sharing at least one region of partial sequence homology, into a suitable host cell;
2) growing the host cell under conditions which promote sequence reorganization resulting in a hybrid polynucleotide in operable linkage;
3) expressing a hybrid polypeptide encoded by the hybrid polynucleotide;
4) screening the hybrid polypeptide under conditions which promote identification of enhanced biological activity; and
5) isolating the a polynucleotide encoding the hybrid polypeptide.

Methods for screening for various enzyme activities are known to those of skill in the art and discussed throughout the present specification. Such methods may be employed when isolating the polypeptides and polynucleotides of the present invention.

As representative examples of expression vectors which may be used there may be mentioned viral particles, baculovirus, phage, plasmids, phagemids, cosmids, fosmids, bacterial artificial chromosomes, viral DNA (e.g. vaccinia, adenovirus, foul pox virus, pseudorabies and derivatives of SV40), P1-based artificial chromosomes, yeast plasmids, yeast artificial chromosomes, and any other vectors specific for specific hosts of interest (such as bacillus, aspergillus and yeast). Thus, for example, the DNA may be included in any one of a variety of expression vectors for expressing a polypeptide. Such vectors include chromosomal, nonchromosomal and synthetic DNA sequences. Large numbers of suitable vectors are known to those of skill in the art, and are commercially available. The following vectors are provided by way of example; Bacterial: pQE vectors (Qiagen), pBluescript plasmids, pNH vectors, (lambda-ZAP vectors (Stratagene); ptrc99a, pKK223-3, pDR540, pRIT2T (Pharmacia); Eukaryotic: pXT1, pSG5 (Stratagene), pSVK3, pBPV, pMSG, pSVLSV40 (Pharmacia). However, any other plasmid or other vector may be used as long as they are replicable and viable in the host. Low copy number or high copy number vectors may be employed with the present invention.

A preferred type of vector for use in the present invention contains an f-factor origin replication. The f-factor (or fertility factor) in *E. coli* is a plasmid which effects high frequency transfer of itself during conjugation and less frequent transfer of the bacterial chromosome itself. A particularly preferred embodiment is to use cloning vectors, referred to as "fosmids" or bacterial artificial chromosome (BAC) vectors. These are derived from *E. coli* f-factor which is able to stably integrate large segments of genomic DNA. When integrated with DNA from a mixed uncultured environmental sample, this makes it possible to achieve large genomic fragments in the form of a stable "environmental DNA library."

Another preferred type of vector for use in the present invention is shuttle vector that is optimized for the expression of genes and gene clusters. Such systems may include but are not limited to shuttling systems that shuttle between *E. coli* and another bacteria such as *Streptomyces*. Another preferred type of vector for use in the present invention is a cosmid vector. Cosmid vectors were originally designed to clone and propagate large segments of genomic DNA. Cloning into cosmid vectors is described in detail in "Molecular Cloning: A laboratory Manual" (**Sambrook** et al, 1989).

The DNA sequence in the expression vector is operatively linked to an appropriate expression control sequence(s) (promoter) to direct RNA synthesis. Particular named bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda P_R, P_L and trp. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Selection of the appropriate vector and promoter is well within the level of ordinary skill in the art. The expression vector also contains a ribosome binding site for translation initiation and a transcription terminator. The vector may also include appropriate sequences for amplifying expression. Promoter regions can be selected from any desired gene using CAT (chloramphenicol transferase) vectors or other vectors with selectable markers.

In addition, the expression vectors preferably contain one or more selectable marker genes to provide a phenotypic trait for selection of transformed host cells such as dihydrofolate reductase or neomycin resistance for eukaryotic cell culture, or such as tetracycline or ampicillin resistance in *E. coli*.

Generally, recombinant expression vectors will include origins of replication and selectable markers permitting transformation of the host cell, e.g., the ampicillin resistance gene of *E. coli* and *S. cerevisiae* TRP1 gene, and a promoter derived from a highly-expressed gene to direct transcription of a downstream structural sequence. Such promoters can be derived from operons encoding glycolytic enzymes such as 3-phosphoglycerate kinase (PGK), -factor, acid phosphatase, or heat shock proteins, among others. The heterologous structural sequence is assembled in appropriate phase with translation initiation and termination sequences, and preferably, a leader sequence capable of directing secretion of translated protein into the periplasmic space or extracellular medium.

The cloning strategy permits expression via both vector driven and endogenous promoters; vector promotion may be important with expression of genes whose endogenous promoter will not function in *E. coli*.

The DNA isolated or derived from microorganisms can preferably be inserted into a vector or a plasmid prior to probing for selected DNA. Such vectors or plasmids are preferably those containing expression regulatory sequences, including promoters, enhancers and the like. Such polynucleotides can be part of a vector and/or a composition and still be isolated, in that such vector or composition is not part of its natural environment. Particularly preferred phage or plasmid and methods for introduction and packaging into them are described in detail in the protocol set forth herein.

The selection of the cloning vector depends upon the approach taken, for example, the vector can be any cloning vector with an adequate capacity to multiply repeated copies of a sequence, or multiple sequences that can be successfully transformed and selected in a host cell. One example of such a vector is described in "Polycos vectors: a system for packaging filamentous phage and phagemid vectors using lambda phage packaging extracts" (**Alting-Mecs and Short**, 1993). Propagation/maintenance can be by an antibiotic resistance carried by the cloning vector. After a period of growth, the naturally abbreviated molecules are recovered and identified by size fractionation on a gel or column, or amplified directly. The cloning vector utilized may contain a selectable gene that is disrupted by the insertion of the lengthy construct. As reductive reassortment progresses, the number of repeated units is reduced and the interrupted gene is again expressed and hence selection for the processed construct can be applied. The vector may be an expression/selection vector which will allow for the selection of an expressed product possessing desirable biologically properties. The insert may be positioned downstream of a functional promotor and the desirable property screened by appropriate means.

In vivo reassortment is focused on "inter-molecular" processes collectively referred to as "recombination" which in bacteria, is generally viewed as a "RecA-dependent" phenomenon. The present invention can rely on recombination processes of a host cell to recombine and re-assort sequences, or the cells' ability to mediate reductive processes to decrease the complexity of quasi-repeated sequences in the cell by deletion. This process of "reductive reassortment" occurs by an "intra-molecular", RecA-independent process.

Therefore, in another aspect of the present invention, novel polynucleotides can be generated by the process of reductive reassortment. The method involves the generation of constructs containing consecutive sequences (original encoding sequences), their insertion into an appropriate vector, and their subsequent introduction into an appropriate host cell. The reassortment of the individual molecular identities occurs by combinatorial processes between the consecutive sequences in the construct possessing regions of homology, or between quasi-repeated units. The reassortment process recombines and/or reduces the complexity and extent of the repeated sequences, and results in the production of novel molecular species. Various treatments may be applied to enhance the rate of reassortment. These could include treatment with ultra-violet light, or DNA damaging chemicals, and/or the use of host cell lines displaying enhanced levels of "genetic instability". Thus the reassortment process may involve homologous recombination or the natural property of quasi-repeated sequences to direct their own evolution.

Repeated or "quasi-repeated" sequences play a role in genetic instability. In the present invention, "quasi-repeats" are repeats that are not restricted to their original unit structure. Quasi-repeated units can be presented as an array of sequences in a construct; consecutive units of similar sequences. Once ligated, the junctions between the consecutive sequences become essentially invisible and the quasi-repetitive nature of the resulting construct is now continuous at the molecular level. The deletion process the cell performs to reduce the complexity of the resulting construct operates between the quasi-repeated sequences. The quasi-repeated units provide a practically limitless repertoire of templates upon which slippage events can occur. The constructs containing the quasi-repeats thus effectively provide sufficient molecular elasticity that deletion (and potentially insertion) events can occur virtually anywhere within the quasi-repetitive units.

When the quasi-repeated sequences are all ligated in the same orientation, for instance head to tail or vice versa, the cell cannot distinguish individual units. Consequently, the reductive process can occur throughout the sequences. In contrast, when for example, the units are presented head to head, rather than head to tail, the inversion delineates the endpoints of the adjacent unit so that deletion formation will favor the loss of discrete units. Thus, it is preferable with the present method that the sequences are in the same orientation. Random orientation of quasi-repeated sequences will result in the loss of reassortment efficiency, while consistent orientation of the sequences will offer the highest efficiency. However, while having fewer of the contiguous sequences in the same orientation decreases the efficiency, it may still provide sufficient elasticity for the effective recovery of novel molecules. Constructs can be made with the quasi-repeated sequences in the same orientation to allow higher efficiency.

Sequences can be assembled in a head to tail orientation using any of a variety of methods, including the following:

a) Primers that include a poly-A head and poly-T tail which when made single-stranded would provide orientation can be utilized. This is accomplished by having the first few bases of the primers made from RNA and hence easily removed RNAseH.

b) Primers that include unique restriction cleavage sites can be utilized. Multiple sites, a battery of unique sequences, and repeated synthesis and ligation steps would be required.

c) The inner few bases of the primer could be thiolated and an exonuclease used to produce properly tailed molecules.

The recovery of the re-assorted sequences relies on the identification of cloning vectors with a reduced RI. The re-assorted encoding sequences can then be recovered by amplification. The products are re-cloned and expressed. The recovery of cloning vectors with reduced RI can be effected by:

1) The use of vectors only stably maintained when the construct is reduced in complexity.

2) The physical recovery of shortened vectors by physical procedures. In this case, the cloning vector would be recovered using standard plasmid isolation

procedures and size fractionated on either an agarose gel, or column with a low molecular weight cut off utilizing standard procedures.

3) The recovery of vectors containing interrupted genes which can be selected when insert size decreases.

4) The use of direct selection techniques with an expression vector and the appropriate selection.

Encoding sequences (for example, genes) from related organisms may demonstrate a high degree of homology and encode quite diverse protein products. These types of sequences are particularly useful in the present invention as quasi-repeats. However, while the examples illustrated below demonstrate the reassortment of nearly identical original encoding sequences (quasi-repeats), this process is not limited to such nearly identical repeats.

The following example demonstrates the method of the invention. Encoding nucleic acid sequences (quasi-repeats) derived from three (3) unique species are depicted. Each sequence encodes a protein with a distinct set of properties. Each of the sequences differs by a single or a few base pairs at a unique position in the sequence which are designated "A", "B" and "C". The quasi-repeated sequences are separately or collectively amplified and ligated into random assemblies such that all possible permutations and combinations are available in the population of ligated molecules. The number of quasi-repeat units can be controlled by the assembly conditions. The average number of quasi-repeated units in a construct is defined as the repetitive index (RI).

Once formed, the constructs may, or may not be size fractionated on an agarose gel according to published protocols, inserted into a cloning vector, and transfected into an appropriate host cell. The cells are then propagated and "reductive reassortment" is effected. The rate of the reductive reassortment process may be stimulated by the introduction of DNA damage if desired. Whether the reduction in RI is mediated by deletion formation between repeated sequences by an "intra-molecular" mechanism, or mediated by recombination-like events through "inter-molecular" mechanisms is immaterial. The end result is a reassortment of the molecules into all possible combinations.

Optionally, the method comprises the additional step of screening the library members of the shuffled pool to identify individual shuffled library members having the ability to bind or otherwise interact (e.g., such as catalytic antibodies) with a predetermined macromolecule, such as for example a proteinaceous receptor, peptide oligosaccharide, viron, or other predetermined compound or structure.

The displayed polypeptides, antibodies, peptidomimetic antibodies, and variable region sequences that are identified from such libraries can be used for therapeutic, diagnostic, research and related purposes (e.g., catalysts, solutes for increasing osmolarity of an aqueous solution, and the like), and/or can be subjected to one or more additional cycles of shuffling and/or affinity selection. The method can be modified such that the step of selecting for a phenotypic characteristic can be other than of binding affinity for a predetermined molecule (e.g., for catalytic activity, stability oxidation resistance, drug resistance, or detectable phenotype conferred upon a host cell).

The present invention provides a method for generating libraries of displayed antibodies suitable for affinity interactions screening. The method comprises (1) obtaining first a plurality of selected library members comprising a displayed antibody and an associated polynucleotide encoding said displayed antibody, and obtaining said associated polynucleotide encoding for said displayed antibody and obtaining said associated polynucleotides or copies thereof, wherein said associated polynucleotides comprise a region of substantially identical variable region framework sequence, and (2) introducing said polynucleotides into a suitable host cell and growing the cells under conditions which promote recombination and reductive reassortment resulting in shuffled polynucleotides. CDR combinations comprised by the shuffled pool are not present in the first plurality of selected library members, said shuffled pool composing a library of displayed antibodies comprising CDR permutations and suitable for affinity interaction screening. Optionally, the shuffled pool is subjected to affinity screening to select shuffled library members which bind to a predetermined epitope (antigen) and thereby selecting a plurality of selected shuffled library members. Further, the plurality of selectively shuffled library members can be shuffled and screened iteratively, from 1 to about 1000 cycles or as desired until library members having a desired binding affinity are obtained.

In another aspect of the invention, it is envisioned that prior to or during recombination or reassortment, polynucleotides generated by the method of the present invention can be subjected to agents or processes which promote the introduction of mutations into the original polynucleotides. The introduction of such mutations would increase the diversity of resulting hybrid polynucleotides and polypeptides encoded therefrom. The agents or processes which promote mutagenesis can include, but are not limited to: (+)-CC-1065, or a synthetic analog such as (+)-CC-1065-(N3-Adenine, see **Sun and Hurley**, 1992); an N-acelylated or deacetylated 4'-fluro-4-aminobiphenyl adduct capable of inhibiting DNA synthesis (see, for example, **van de Poll** et al, 1992); or a N-acetylated or deacetylated 4-aminobiphenyl adduct capable of inhibiting DNA synthesis (see also, **van de Poll** et al, 1992, pp. 751-758); trivalent chromium, a trivalent chromium salt, a polycyclic aromatic hydrocarbon ("PAH") DNA adduct capable of inhibiting DNA replication, such as 7-bromomethyl-benz[*a*]anthracene ("BMA"), tris(2,3-dibromopropyl)phosphate ("Tris-BP"), 1,2-dibromo-3-chloropropane ("DBCP"), 2-bromoacrolein (2BA), benzo[*a*]pyrene-7,8-dihydrodiol-9-10-epoxide ("BPDE"), a platinum(II) halogen salt, N-hydroxy-2-amino-3-methylimidazo[4,5-*f*]-quinoline ("N-hydroxy-IQ"), and N-hydroxy-2-amino-1-methyl-6-phenylimidazo[4,5-*f*]-pyridine ("N-hydroxy-PhIP"). Especially preferred "means for slowing or halting PCR amplification consist of UV light (+)-CC-1065 and (+)-CC-1065-(N3-Adenine). Particularly encompassed means are DNA adducts or polynucleotides comprising the DNA adducts from the polynucleotides or polynucleotides pool, which can be released or removed by a process including heating the solution comprising the polynucleotides prior to further processing.

In another aspect, this invention provides for using UV light to mutagenize polynucleotides. One use of such a technique is as follows: one microgram samples of template DNA are obtained and treated with U.V. light to cause the formation of dimers, including TT dimers, particularly purine dimers. U.V. exposure is limited so that only a few photoproducts are generated per gene on the template DNA sample. Multiple samples are treated with U.V. light for varying periods of time to obtain template DNA samples with varying numbers of dimers from U.V. exposure. A random priming kit which utilizes a non-proofreading polymease (for example, Prime-It II Random Primer Labeling kit by Stratagene Cloning Systems) is utilized to generate different size polynucleotides by priming at random sites on templates which are prepared by U.V. light (as described above) and extending along the templates. The priming protocols such as described in the Prime-It II Random Primer Labeling kit may be utilized to extend the primers. The dimers formed by U.V. exposure serve as a roadblock for the extension by the non-proofreading polymerase. Thus, a pool of random size polynucleotides is present after extension with the random primers is finished.

In another aspect the present invention is directed to a method of producing recombinant proteins having biological activity by treating a sample comprising double-stranded template polynucleotides encoding a wild-type protein under conditions according to the present invention which provide for the production of hybrid or re-assorted polynucleotides.

The invention also provides the use of polynucleotide shuffling to shuffle a population of viral genes (e.g., capsid proteins, spike glycoproteins, polymerases, and proteases) or viral genomes (e.g., paramyxoviridae, orthomyxoviridae, herpesviruses, retroviruses, reoviruses and rhinoviruses). In an embodiment, the invention provides a method for shuffling sequences encoding all or portions of immunogenic viral proteins to generate novel combinations of epitopes as well as novel epitopes created by recombination; such shuffled viral proteins may comprise epitopes or combinations of epitopes as well as novel epitopes created by recombination; such shuffled viral proteins may comprise epitopes or combinations of epitopes which are likely to arise in the natural environment as a consequence of viral evolution; (e.g., such as recombination of influenza virus strains).

The invention also provides a method suitable for shuffling polynucleotide sequences for generating gene therapy vectors and replication-defective gene therapy constructs, such as may be used for human gene therapy, including but not limited to vaccination vectors for DNA-based vaccination, as well as anti-neoplastic gene therapy and other general therapy formats.

In the polypeptide notation used herein, the left-hand direction is the amino terminal direction and the right-hand direction is the carboxy-terminal direction, in accordance with standard usage and convention. Similarly, unless specified otherwise, the left-hand end of single-stranded polynucleotide sequences is the 5' end; the left-hand direction of double-stranded polynucleotide sequences is referred to as the 5' direction. The direction of 5' to 3' addition of nascent RNA transcripts is referred to as the transcription direction; sequence regions on the DNA strand having the same sequence as the RNA and which are 5' to the 5' end of the RNA transcript are referred to as "upstream sequences"; sequence regions on the DNA strand having the same sequence as the RNA and which are 3' to the 3' end of the coding RNA transcript are referred to as "downstream sequences".

3.1. SATURATION MUTAGENESIS

In one aspect, this invention provides for the use of proprietary codon primers (containing a degenerate N,N,G/T sequence) to introduce point mutations into a polynucleotide, so as to generate a set of progeny polypeptides in which a full range of single amino acid substitutions is represented at each amino acid position. The oligos used are comprised contiguously of a first homologous sequence, a degenerate N,N,G/T sequence, and preferably but not necessarily a second homologous sequence. The downstream progeny translational products from the use of such oligos include all possible amino acid changes at each amino acid site along the polypeptide, because the degeneracy of the N,N,G/T sequence includes codons for all 20 amino acids.

In one aspect, one such degenerate oligo (comprised of one degenerate N,N,G/T cassette) is used for subjecting each original codon in a parental polynucleotide template to a full range of codon substitutions. In another aspect, at least two degenerate N,N,G/T cassettes are used – either in the same oligo or not, for subjecting at least two original codons in a parental polynucleotide template to a full range of codon substitutions. Thus, more than one N,N,G/T sequence can be contained in one oligo to introduce amino acid mutations at more than one site. This plurality of N,N,G/T sequences can be directly contiguous, or separated by one or more additional nucleotide sequence(s). In another aspect, oligos serviceable for introducing additions and deletions can be used either alone or in combination with the codons containing an N,N,G/T sequence, to introduce any combination or permutation of amino acid additions, deletions, and/or substitutions.

In a particular exemplification, it is possible to simultaneously mutagenize two or more contiguous amino acid positions using an oligo that contains contiguous N,N,G/T triplets, i.e. a degenerate $(N,N,G/T)_n$ sequence.

In another aspect, the present invention provides for the use of degenerate cassettes having less degeneracy than the N,N,G/T sequence. For example, it may be desirable in some instances to use (e.g. in an oligo) a degenerate triplet sequence comprised of only one N, where said N can be in the first second or third position of the triplet. Any other bases including any combinations and permutations thereof can be used in the remaining two positions of the triplet. Alternatively, it may be desirable in some instances to use (e.g. in an oligo) a degenerate N,N,N triplet sequence, or an N,N, G/C triplet sequence.

It is appreciated, however, that the use of a degenerate triplet (such as N,N,G/T or an N,N, G/C triplet sequence) as disclosed in the instant invention is advantageous for several reasons. In one aspect, this invention provides a means to systematically and fairly easily generate the substitution of the full range of possible amino acids (for a total of 20 amino acids) into each and every amino acid position in a polypeptide. Thus, for a 100 amino acid polypeptide, the instant invention provides a way to systematically and fairly easily generate 2000 distinct species (i.e. 20 possible amino acids per position X 100 amino acid positions). It is appreciated that there is provided, through the use of an oligo containing a degenerate N,N,G/T or an N,N, G/C triplet sequence, 32 individual sequences that code for 20 possible amino acids. Thus, in a reaction vessel in which a parental polynucleotide sequence is subjected to saturation mutagenesis using one such oligo, there are generated 32 distinct progeny polynucleotides encoding 20 distinct polypeptides. In contrast, the use of a non-degenerate oligo in site-directed mutagenesis leads to only one progeny polypeptide product per reaction vessel.

This invention also provides for the use of nondegenerate oligos, which can optionally be used in combination with degenerate primers disclosed. It is appreciated that in some situations, it is advantageous to use nondegenerate oligos to generate specific point mutations in a working polynucleotide. This provides a means to generate specific silent point mutations, point mutations leading to corresponding amino acid changes, and point mutations that cause the generation of stop codons and the corresponding expression of polypeptide fragments.

Thus, in a preferred embodiment of this invention, each saturation mutagenesis reaction vessel contains polynucleotides encoding at least 20 progeny polypeptide molecules such that all 20 amino acids are represented at the one specific amino acid position corresponding to the codon position mutagenized in the parental polynucleotide. The 32-fold degenerate progeny polypeptides generated from each saturation mutagenesis reaction vessel can be subjected to clonal amplification (e.g. cloned into a suitable *E. coli* host using an expression vector) and subjected to expression screening. When an individual progeny polypeptide is identified by screening to display a favorable change in property (when compared to the parental polypeptide), it can be sequenced to identify the correspondingly favorable amino acid substitution contained therein.

It is appreciated that upon mutagenizing each and every amino acid position in a parental polypeptide using saturation mutagenesis as disclosed herein, favorable amino acid changes may be identified at more than one amino acid position. One or more new progeny molecules can be generated that contain a combination of all or part of these favorable amino acid substitutions. For example, if 2 specific favorable amino acid changes are identified in each of 3 amino acid positions in a polypeptide, the permutations include 3 possibilities at each position (no change from the original amino acid, and each of two favorable changes) and 3 positions. Thus, there are 3 x 3 x 3 or 27 total possibilities, including 7 that were previously examined - 6 single point mutations (i.e. 2 at each of three positions) and no change at any position.

In yet another aspect, site-saturation mutagenesis can be used together with shuffling, chimerization, recombination and other mutagenizing processes, along with screening. This invention provides for the use of any mutagenizing process(es), including saturation mutagenesis, in an iterative manner. In one exemplification, the iterative use of any mutagenizing process(es) is used in combination with screening.

Thus, in a non-limiting exemplification, this invention provides for the use of saturation mutagenesis in combination with additional mutagenization processes, such as process where two or more related polynucleotides are introduced into a suitable host cell such that a hybrid polynucleotide is generated by recombination and reductive reassortment.

In addition to performing mutagenesis along the entire sequence of a gene, the instant invention provides that mutagenesis can be use to replace each of any number of bases in a polynucleotide sequence, wherein the number of bases to be mutagenized is preferably every integer from 15 to 100,000. Thus, instead of mutagenizing every position along a molecule, one can subject every a discrete number of bases (preferably a subset totaling from 15 to 100,000) to mutagenesis. Preferably, a separate nucleotide is used for mutagenizing each position or group of positions along a polynucleotide sequence. A group of 3 positions to be mutagenized may be a codon. The mutations are preferably introduced using a mutagenic primer, containing a heterologous cassette, also referred to as a mutagenic cassette. Preferred cassettes can have from 1 to 500 bases. Each nucleotide position in such heterologous cassettes be N, A, C, G, T, A/C, A/G, A/T, C/G, C/T, G/T, C/G/T, A/G/T, A/C/T, A/C/G, or E, where E is any base that is not A, C, G, or T (E can be referred to as a designer oligo). The tables below show exemplary tri-nucleotide cassettes (there are over 3000 possibilities in addition to N,N,G/T and N,N,N and N,N,A/C).

In a general sense, saturation mutagenesis is comprised of mutagenizing a complete set of mutagenic cassettes (wherein each cassette is preferably 1-500 bases in length) in defined polynucleotide sequence to be mutagenized (wherein the sequence to be mutagenized is preferably from 15 to 100,000 bases in length). Thusly, a group of mutations (ranging from 1 to 100 mutations) is introduced into each cassette to be mutagenized. A grouping of mutations to be introduced into one cassette can be different or the same from a second grouping of mutations to be introduced into a second cassette during the application of one round of saturation mutagenesis. Such groupings are exemplified by deletions, additions, groupings of particular codons, and groupings of particular nucleotide cassettes.

Defined sequences to be mutagenized (see Fig. 20) include preferably a whole gene, pathway, cDNA, an entire open reading frame (ORF), and entire promoter, enhancer, repressor/transactivator, origin of replication, intron, operator, or any polynucleotide functional group. Generally, a preferred "defined sequences" for this purpose may be any polynucleotide that a 15 base-polynucleotide sequence, and polynucleotide sequences of lengths between 15 bases and 15,000 bases (this invention specifically names every integer in between). Considerations in choosing groupings of codons include types of amino acids encoded by a degenerate mutagenic cassette.

In a particularly preferred exemplification a grouping of mutations that can be introduced into a mutagenic cassette (see Tables 1-85), this invention specifically provides for degenerate codon substitutions (using degenerate oligos) that code for 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, and 20 amino acids at each position, and a library of polypeptides encoded thereby.

3.2. CHIMERIZATIONS

3.2.1 "SHUFFLING"

Nucleic acid shuffling is a method for *in vitro* or *in vivo* homologous recombination of pools of shorter or smaller polynucleotides to produce a polynucleotide or polynucleotides. Mixtures of related nucleic acid sequences or polynucleotides are subjected to sexual PCR to provide random polynucleotides, and reassembled to yield a library or mixed population of recombinant hybrid nucleic acid molecules or polynucleotides.

In contrast to cassette mutagenesis, only shuffling and error-prone PCR allow one to mutate a pool of sequences blindly (without sequence information other than primers).

The advantage of the mutagenic shuffling of this invention over error-prone PCR alone for repeated selection can best be explained with an example from antibody engineering. Consider DNA shuffling as compared with error-prone PCR (not sexual PCR). The initial library of selected pooled sequences can consist of related sequences of diverse origin (i.e. antibodies from naive mRNA) or can be derived by any type of mutagenesis (including shuffling) of a single antibody gene. A collection of selected complementarity determining regions ("CDRs") is obtained after the first round of affinity selection. In the diagram the thick CDRs confer onto the antibody molecule increased affinity for the antigen. Shuffling allows the free combinatorial association of all of the CDR1s with all of the CDR2s with all of the CDR3s, for example.

This method differs from error-prone PCR, in that it is an inverse chain reaction. In error-prone PCR, the number of polymerase start sites and the number of molecules grows exponentially. However, the sequence of the polymerase start sites and the sequence of the molecules remains essentially the same. In contrast, in nucleic acid reassembly or shuffling of random polynucleotides the number of start sites and the number (but not size) of the random polynucleotides decreases over time. For polynucleotides derived from whole plasmids the theoretical endpoint is a single, large concatemeric molecule.

Since cross-overs occur at regions of homology, recombination will primarily occur between members of the same sequence family. This discourages combinations of CDRs that are grossly incompatible (e.g., directed against different epitopes of the same antigen). It is contemplated that multiple families of sequences can be shuffled in the same reaction. Further, shuffling generally conserves the relative order, such that, for example, CDR1 will not be found in the position of CDR2.

Rare shufflants will contain a large number of the best (eg. highest affinity) CDRs and these rare shufflants may be selected based on their superior affinity.

CDRs from a pool of 100 different selected antibody sequences can be permutated in up to 1006 different ways. This large number of permutations cannot be represented in a single library of DNA sequences. Accordingly, it is contemplated that multiple cycles of DNA shuffling and selection may be required depending on the length of the sequence and the sequence diversity desired.

Error-prone PCR, in contrast, keeps all the selected CDRs in the same relative sequence, generating a much smaller mutant cloud.

The template polynucleotide which may be used in the methods of this invention may be DNA or RNA. It may be of various lengths depending on the size of the gene or shorter or smaller polynucleotide to be recombined or reassembled. Preferably, the template polynucleotide is from 50 bp to 50 kb. It is contemplated that entire vectors containing the nucleic acid encoding the protein of interest can be used in the methods of this invention, and in fact have been successfully used.

The template polynucleotide may be obtained by amplification using the PCR reaction (USPN 4,683,202 and USPN 4,683,195) or other amplification or cloning methods. However, the removal of free primers from the PCR products before subjecting them to pooling of the PCR products and sexual PCR may provide more efficient results. Failure to adequately remove the primers from the original pool before sexual PCR can lead to a low frequency of crossover clones.

The template polynucleotide often should be double-stranded. A double-stranded nucleic acid molecule is recommended to ensure that regions of the resulting single-stranded polynucleotides are complementary to each other and thus can hybridize to form a double-stranded molecule.

It is contemplated that single-stranded or double-stranded nucleic acid polynucleotides having regions of identity to the template polynucleotide and regions of heterology to the template polynucleotide may be added to the template polynucleotide, at this step. It is also contemplated that two different but related polynucleotide templates can be mixed at this step.

The double-stranded polynucleotide template and any added double-or single-stranded polynucleotides are subjected to sexual PCR which includes slowing or halting to provide a mixture of from about 5 bp to 5 kb or more. Preferably the size of the random polynucleotides is from about 10 bp to 1000 bp, more preferably the size of the polynucleotides is from about 20 bp to 500 bp.

Alternatively, it is also contemplated that double-stranded nucleic acid having multiple nicks may be used in the methods of this invention. A nick is a break in one strand of the double-stranded nucleic acid. The distance between such nicks is preferably 5 bp to 5 kb, more preferably between 10 bp to 1000 bp. This can provide areas of self-priming to produce shorter or smaller polynucleotides to be included with the polynucleotides resulting from random primers, for example.

The concentration of any one specific polynucleotide will not be greater than 1% by weight of the total polynucleotides, more preferably the concentration of any one specific nucleic acid sequence will not be greater than 0.1% by weight of the total nucleic acid.

The number of different specific polynucleotides in the mixture will be at least about 100, preferably at least about 500, and more preferably at least about 1000.

At this step single-stranded or double-stranded polynucleotides, either synthetic or natural, may be added to the random double-stranded shorter or smaller polynucleotides in order to increase the heterogeneity of the mixture of polynucleotides.

It is also contemplated that populations of double-stranded randomly broken polynucleotides may be mixed or combined at this step with the polynucleotides from the sexual PCR process and optionally subjected to one or more additional sexual PCR cycles.

Where insertion of mutations into the template polynucleotide is desired, single-stranded or double-stranded polynucleotides having a region of identity to the template polynucleotide and a region of heterology to the template polynucleotide may be added in a 20 fold excess by weight as compared to the total nucleic acid, more preferably the single-stranded polynucleotides may be added in a 10 fold excess by weight as compared to the total nucleic acid.

Where a mixture of different but related template polynucleotides is desired, populations of polynucleotides from each of the templates may be combined at a ratio of less than about 1:100, more preferably the ratio is less than about 1:40. For example, a backcross of the wild-type polynucleotide with a population of mutated polynucleotide may be desired to eliminate neutral mutations (e.g., mutations yielding an insubstantial alteration in the phenotypic property being selected for). In such an example, the ratio of randomly provided wild-type polynucleotides which may be added to the randomly provided sexual PCR cycle hybrid polynucleotides is approximately 1:1 to about 100:1, and more preferably from 1:1 to 40:1.

The mixed population of random polynucleotides are denatured to form single-stranded polynucleotides and then re-annealed. Only those single-stranded polynucleotides having regions of homology with other single-stranded polynucleotides will re-anneal.

The random polynucleotides may be denatured by heating. One skilled in the art could determine the conditions necessary to completely denature the double-stranded nucleic acid. Preferably the temperature is from 80°C to 100°C, more preferably the temperature is from 90°C to 96°C. other methods which may be used to denature the polynucleotides include pressure (36) and pH.

The polynucleotides may be re-annealed by cooling. Preferably the temperature is from 20°C to 75°C, more preferably the temperature is from 40°C to 65°C. If a high frequency of crossovers is needed based on an average of only 4 consecutive bases of homology, recombination can be forced by using a low annealing temperature, although the process becomes more difficult. The degree of renaturation which occurs will depend on the degree of homology between the population of single-stranded polynucleotides.

Renaturation can be accelerated by the addition of polyethylene glycol ("PEG") or salt. The salt concentration is preferably from 0 mM to 200 mM, more preferably the salt concentration is from 10 mM to 100 mm. The salt may be KCl or NaCl. The concentration of PEG is preferably from 0% to 20%, more preferably from 5% to 10%.

The annealed polynucleotides are next incubated in the presence of a nucleic acid polymerase and dNTP's (i.e. dATP, dCTP, DGTP and dTTP). The nucleic acid polymerase may be the Klenow fragment, the Taq polymerase or any other DNA polymerase known in the art.

The approach to be used for the assembly depends on the minimum degree of homology that should still yield crossovers. If the areas of identity are large, Taq polymerase can be used with an annealing temperature of between 45-65°C. If the areas of identity are small, Klenow polymerase can be used with an annealing temperature of between 20-30°C. One skilled in the art could vary the temperature of annealing to increase the number of cross-overs achieved.

The polymerase may be added to the random polynucleotides prior to annealing, simultaneously with annealing or after annealing.

The cycle of denaturation, renaturation and incubation in the presence of polymerase is referred to herein as shuffling or reassembly of the nucleic acid. This cycle is repeated for a desired number of times. Preferably the cycle is repeated from 2 to 50 times, more preferably the sequence is repeated from 10 to 40 times.

The resulting nucleic acid is a larger double-stranded polynucleotide of from about 50 bp to about 100 kb, preferably the larger polynucleotide is from 500 bp to 50 kb.

This larger polynucleotide may contain a number of copies of a polynucleotide having the same size as the template polynucleotide in tandem. This concatemeric polynucleotide is then denatured into single copies of the template polynucleotide. The result will be a population of polynucleotides of approximately the same size as the template polynucleotide. The population will be a mixed population where single or double-stranded polynucleotides having an area of identity and an area of heterology have been added to the template polynucleotide prior to shuffling. These polynucleotides are then cloned into the appropriate vector and the ligation mixture used to transform bacteria.

It is contemplated that the single polynucleotides may be obtained from the larger concatemeric polynucleotide by amplification of the single polynucleotide prior to cloning by a variety of methods including PCR (**USPN 4,683,195 and USPN 4,683,202**), rather than by digestion of the concatemer.

The vector used for cloning is not critical provided that it will accept a polynucleotide of the desired size. If expression of the particular polynucleotide is desired, the cloning vehicle should further comprise transcription and translation signals next to the site of insertion of the polynucleotide to allow expression of the polynucleotide in the host cell. Preferred vectors include the pUC series and the pBR series of plasmids.

The resulting bacterial population will include a number of recombinant polynucleotides having random mutations. This mixed population may be tested to identify the desired recombinant polynucleotides. The method of selection will depend on the polynucleotide desired.

For example, if a polynucleotide which encodes a protein with increased binding efficiency to a ligand is desired, the proteins expressed by each of the portions of the polynucleotides in the population or library may be tested for their ability to bind to the ligand by methods known in the art (i.e. panning, affinity chromatography). If a polynucleotide which encodes for a protein with increased drug resistance is desired, the proteins expressed by each of the polynucleotides in the population or library may be tested for their ability to confer drug resistance to the host organism. One skilled in the art, given knowledge of the desired protein, could readily test the population to identify polynucleotides which confer the desired properties onto the protein.

It is contemplated that one skilled in the art could use a phage display system in which fragments of the protein are expressed as fusion proteins on the phage surface (Pharmacia, Milwaukee WI). The recombinant DNA molecules are cloned into the phage DNA at a site which results in the transcription of a fusion protein a portion of which is encoded by the recombinant DNA molecule. The phage containing the recombinant nucleic acid molecule undergoes replication and transcription in the cell. The leader sequence of the fusion protein directs the transport of the fusion protein to the tip of the phage particle. Thus the fusion protein which is partially encoded by the recombinant DNA molecule is displayed on the phage particle for detection and selection by the methods described above.

It is further contemplated that a number of cycles of nucleic acid shuffling may be conducted with polynucleotides from a sub-population of the first population, which sub-population contains DNA encoding the desired recombinant protein. In this manner, proteins with even higher binding affinities or enzymatic activity could be achieved.

It is also contemplated that a number of cycles of nucleic acid shuffling may be conducted with a mixture of wild-type polynucleotides and a sub-population of nucleic acid from the first or subsequent rounds of nucleic acid shuffling in order to remove any silent mutations from the sub-population.

Any source of nucleic acid, in purified form can be utilized as the starting nucleic acid. Thus the process may employ DNA or RNA including messenger RNA, which DNA or RNA may be single or double stranded. In addition, a DNA-RNA hybrid which contains one strand of each may be utilized. The nucleic acid sequence may be of various lengths depending on the size of the nucleic acid sequence to be mutated. Preferably the specific nucleic acid sequence is from 50 to 50000 base pairs. It is contemplated that entire vectors containing the nucleic acid encoding the protein of interest may be used in the methods of this invention.

The nucleic acid may be obtained from any source, for example, from plasmids such a pBR322, from cloned DNA or RNA or from natural DNA or RNA from any source including bacteria, yeast, viruses and higher organisms such as plants or animals. DNA or RNA may be extracted from blood or tissue material. The template polynucleotide may be obtained by amplification using the polynucleotide chain reaction (PCR, see **USPN 4,683,202** and **USPN 4,683,195**). Alternatively, the polynucleotide may be present in a vector present in a cell and sufficient nucleic acid may be obtained by culturing the cell and extracting the nucleic acid from the cell by methods known in the art.

Any specific nucleic acid sequence can be used to produce the population of hybrids by the present process. It is only necessary that a small population of hybrid sequences of the specific nucleic acid sequence exist or be created prior to the present process.

The initial small population of the specific nucleic acid sequences having mutations may be created by a number of different methods. Mutations may be created by error-prone PCR. Error-prone PCR uses low-fidelity polymerization conditions to introduce a low level of point mutations randomly over a long sequence. Alternatively, mutations can be introduced into the template polynucleotide by oligonucleotide-directed mutagenesis. In oligonucleotide-directed mutagenesis, a short sequence of the polynucleotide is removed from the polynucleotide using restriction enzyme digestion and is replaced with a synthetic polynucleotide in which various bases have been altered from the original sequence. The polynucleotide sequence can also be altered by chemical mutagenesis. Chemical mutagens include, for example, sodium bisulfite, nitrous acid, hydroxylamine, hydrazine or formic acid. Other agents which are analogues of nucleotide precursors include nitrosoguanidine, 5-bromouracil, 2-aminopurine, or acridine. Generally, these agents are added to the PCR reaction in place of the nucleotide precursor thereby mutating the sequence. Intercalating agents such as proflavine, acriflavine, quinacrine and the like can also be used. Random mutagenesis of the polynucleotide sequence can also be achieved by irradiation with X-rays or ultraviolet light. Generally, plasmid polynucleotides so mutagenized are introduced into *E. coli* and propagated as a pool or library of hybrid plasmids.

Alternatively the small mixed population of specific nucleic acids may be found in nature in that they may consist of different alleles of the same gene or the same gene from different related species (i.e., cognate genes). Alternatively, they may be related DNA sequences found within one species, for example, the immunoglobulin genes.

Once the mixed population of the specific nucleic acid sequences is generated, the polynucleotides can be used directly or inserted into an appropriate cloning vector, using techniques well-known in the art.

The choice of vector depends on the size of the polynucleotide sequence and the host cell to be employed in the methods of this invention. The templates of this invention may be plasmids, phages, cosmids, phagemids, viruses (e.g., retroviruses, parainfluenzavirus, herpesviruses, reoviruses, paramyxoviruses, and the like), or selected portions thereof (e.g., coat protein, spike glycoprotein, capsid protein). For example, cosmids and phagemids are preferred where the specific nucleic acid sequence to be mutated is larger because these vectors are able to stably propagate large polynucleotides.

If the mixed population of the specific nucleic acid sequence is cloned into a vector it can be clonally amplified by inserting each vector into a host cell and allowing the host cell to amplify the vector. This is referred to as clonal amplification because while the absolute number of nucleic acid sequences increases, the number of hybrids does not increase. Utility can be readily determined by screening expressed polypeptides.

The DNA shuffling method of this invention can be performed blindly on a pool of unknown sequences. By adding to the reassembly mixture oligonucleotides (with ends that are homologous to the sequences being reassembled) any sequence mixture can be incorporated at any specific position into another sequence mixture. Thus, it is contemplated that mixtures of synthetic oligonucleotides, PCR polynucleotides or even whole genes can be mixed into another sequence library at defined positions. The insertion of one sequence (mixture) is independent from the insertion of a sequence in another part of the template. Thus, the degree of recombination, the homology required, and the diversity of the library can be independently and simultaneously varied along the length of the reassembled DNA.

This approach of mixing two genes may be useful for the humanization of antibodies from murine hybridomas. The approach of mixing two genes or inserting alternative sequences into genes may be useful for any therapeutically used protein, for example, interleukin I, antibodies, tPA and growth hormone. The approach may also be useful in any nucleic acid for example, promoters or introns or 3' untranslated region or 5' untranslated regions of genes to increase expression or alter specificity of expression of proteins. The approach may also be used to mutate ribozymes or aptamers.

Shuffling requires the presence of homologous regions separating regions of diversity. Scaffold-like protein structures may be particularly suitable for shuffling. The conserved scaffold determines the overall folding by self-association, while displaying relatively unrestricted loops that mediate the specific binding. Examples of such scaffolds are the immunoglobulin beta-barrel, and the four-helix bundle which are well-known in the art. This shuffling can be used to create scaffold-like proteins with various combinations of mutated sequences for binding.

3.2.1.1. *In vitro* Shuffling

The equivalents of some standard genetic matings may also be performed by shuffling *in vitro*. For example, a "molecular backcross" can be performed by repeatedly mixing the hybrid's nucleic acid with the wild-type nucleic acid while selecting for the mutations of interest. As in traditional breeding, this approach can be used to combine phenotypes from different sources into a background of choice. It is useful, for example, for the removal of neutral mutations that affect unselected characteristics (i.e. immunogenicity). Thus it can be useful to determine which mutations in a protein are involved in the enhanced biological activity and which are not, an advantage which cannot be achieved by error-prone mutagenesis or cassette mutagenesis methods.

Large, functional genes can be assembled correctly from a mixture of small random polynucleotides. This reaction may be of use for the reassembly of genes from the highly fragmented DNA of fossils. In addition random nucleic acid fragments from fossils may be combined with polynucleotides from similar genes from related species.

It is also contemplated that the method of this invention can be used for the *in vitro* amplification of a whole genome from a single cell as is needed for a variety of research and diagnostic applications. DNA amplification by PCR is in practice limited to a length of about 40 kb. Amplification of a whole genome such as that of *E.* coli (5, 000 kb) by PCR would require about 250 primers yielding 125 forty kb polynucleotides. This approach is not practical due to the unavailability of sufficient sequence data. On the other hand, random production of polynucleotides of the genome with sexual PCR cycles, followed by gel purification of small polynucleotides will provide a multitude of possible primers. Use of this mix of random small polynucleotides as primers in a PCR reaction alone or with the whole genome as the template should result in an inverse chain reaction with the theoretical endpoint of a single concatamer containing many copies of the genome.

100 fold amplification in the copy number and an average polynucleotide size of greater than 50 kb may be obtained when only random polynucleotides are used. It is thought that the larger concatamer is generated by overlap of many smaller polynucleotides. The quality of specific PCR products obtained using synthetic primers will be indistinguishable from the product obtained from unamplified DNA. It is expected that this approach will be useful for the mapping of genomes.

The polynucleotide to be shuffled can be produced as random or non-random polynucleotides, at the discretion of the practitioner. Moreover, this invention provides a method of shuffling that is applicable to a wide range of polynucleotide sizes and types, including the step of generating polynucleotide monomers to be used as building blocks in the reassembly of a larger polynucleotide. For example, the building blocks can be fragments of genes or they can be comprised of entire genes or gene pathways, or any combination thereof.

3.2.1.2. *In vivo* Shuffling

In an embodiment of *in vivo* shuffling, the mixed population of the specific nucleic acid sequence is introduced into bacterial or eukaryotic cells under conditions such that at least two different nucleic acid sequences are present in each host cell. The polynucleotides can be introduced into the host cells by a variety of different methods. The host cells can be transformed with the smaller polynucleotides using methods known in the art, for example treatment with calcium chloride. If the polynucleotides are inserted into a phage genome, the host cell can be transfected with the recombinant phage genome having the specific nucleic acid sequences. Alternatively, the nucleic acid sequences can be introduced into the host cell using electroporation, transfection, lipofection, biolistics, conjugation, and the like.

In general, in this embodiment, the specific nucleic acids sequences will be present in vectors which are capable of stably replicating the sequence in the host cell. In addition, it is contemplated that the vectors will encode a marker gene such that host cells having the vector can be selected. This ensures that the mutated specific nucleic acid sequence can be recovered after introduction into the host cell. However, it is contemplated that the entire mixed population of the specific nucleic acid sequences need not be present on a vector sequence. Rather only a sufficient number of sequences need be cloned into vectors to ensure that after introduction of the polynucleotides into the host cells each host cell contains one vector having at least one specific nucleic acid sequence present therein. It is also contemplated that rather than having a subset of the population of the specific nucleic acids sequences cloned into vectors, this subset may be already stably integrated into the host cell.

It has been found that when two polynucleotides which have regions of identity are inserted into the host cells homologous recombination occurs between the two polynucleotides. Such recombination between the two mutated specific nucleic acid sequences will result in the production of double or triple hybrids in some situations.

It has also been found that the frequency of recombination is increased if some of the mutated specific nucleic acid sequences are present on linear nucleic acid molecules. Therefore, in a preferred embodiment, some of the specific nucleic acid sequences are present on linear polynucleotides.

After transformation, the host cell transformants are placed under selection to identify those host cell transformants which contain mutated specific nucleic acid sequences having the qualities desired. For example, if increased resistance to a particular drug is desired then the transformed host cells may be subjected to increased concentrations of the particular drug and those transformants producing mutated proteins able to confer increased drug resistance will be selected. If the enhanced ability of a particular protein to bind to a receptor is desired, then expression of the protein can be induced from the transformants and the resulting protein assayed in a ligand binding assay by methods known in the art to identify that subset of the mutated population which shows enhanced binding to the ligand. Alternatively, the protein can be expressed in another system to ensure proper processing.

Once a subset of the first recombined specific nucleic acid sequences (daughter sequences) having the desired characteristics are identified, they are then subject to a second round of recombination.

In the second cycle of recombination, the recombined specific nucleic acid sequences may be mixed with the original mutated specific nucleic acid sequences (parent sequences) and the cycle repeated as described above. In this way a set of second recombined specific nucleic acids sequences can be identified which have enhanced characteristics or encode for proteins having enhanced properties. This cycle can be repeated a number of times as desired.

It is also contemplated that in the second or subsequent recombination cycle, a backcross can be performed. A molecular backcross can be performed by mixing the desired specific nucleic acid sequences with a large number of the wild-type sequence, such that at least one wild-type nucleic acid sequence and a mutated nucleic acid sequence are present in the same host cell after transformation. Recombination with the wild-type specific nucleic acid sequence will eliminate those neutral mutations that may affect unselected characteristics such as immunogenicity but not the selected characteristics.

In another embodiment of this invention, it is contemplated that during the first round a subset of the specific nucleic acid sequences can be generated as smaller polynucleotides by slowing or halting their PCR amplification prior to introduction into the host cell. The size of the polynucleotides must be large enough to contain some regions of identity with the other sequences so as to homologously recombine with the other sequences. The size of the polynucleotides will range from 0.03 kb to 100 kb more preferably from 0. 2 kb to 10 kb. It is also contemplated that in subsequent rounds, all of the specific nucleic acid sequences other than the sequences selected from the previous round may be utilized to generate PCR polynucleotides prior to introduction into the host cells.

The shorter polynucleotide sequences can be single-stranded or double-stranded. If the sequences were originally single-stranded and have become double-stranded they can be denatured with heat, chemicals or enzymes prior to insertion into the host cell. The reaction conditions suitable for separating the strands of nucleic acid are well known in the art.

The steps of this process can be repeated indefinitely, being limited only by the number of possible hybrids which can be achieved. After a certain number of cycles, all possible hybrids will have been achieved and further cycles are redundant.

In an embodiment the same mutated template nucleic acid is repeatedly recombined and the resulting recombinants selected for the desired characteristic.

Therefore, the initial pool or population of mutated template nucleic acid is cloned into a vector capable of replicating in a bacteria such as *E. coli.* The particular vector is not essential, so long as it is capable of autonomous replication in E. coli. In a preferred embodiment, the vector is designed to allow the expression and production of any protein encoded by the mutated specific nucleic acid linked to the vector. It is also preferred that the vector contain a gene encoding for a selectable marker.

The population of vectors containing the pool of mutated nucleic acid sequences is introduced into the *E. coli* host cells. The vector nucleic acid sequences may be introduced by transformation, transfection or infection in the case of phage. The concentration of vectors used to transform the bacteria is such that a number of vectors is introduced into each cell. Once present in the cell, the efficiency of homologous recombination is such that homologous recombination occurs between the various vectors. This results in the generation of hybrids (daughters) having a combination of mutations which differ from the original parent mutated sequences.

The host cells are then clonally replicated and selected for the marker gene present on the vector. Only those cells having a plasmid will grow under the selection.

The host cells which contain a vector are then tested for the presence of favorable mutations. Such testing may consist of placing the cells under selective pressure, for example, if the gene to be selected is an improved drug resistance gene. If the vector allows expression of the protein encoded by the mutated nucleic acid sequence, then such selection may include allowing expression of the protein so encoded, isolation of the protein and testing of the protein to determine whether, for example, it binds with increased efficiency to the ligand of interest.

Once a particular daughter mutated nucleic acid sequence has been identified which confers the desired characteristics, the nucleic acid is isolated either already linked to the vector or separated from the vector. This nucleic acid is then mixed with the first or parent population of nucleic acids and the cycle is repeated.

It has been shown that by this method nucleic acid sequences having enhanced desired properties can be selected.

In an alternate embodiment, the first generation of hybrids are retained in the cells and the parental mutated sequences are added again to the cells. Accordingly, the first cycle of Embodiment I is conducted as described above. However, after the daughter nucleic acid sequences are identified, the host cells containing these sequences are retained.

The parent mutated specific nucleic acid population, either as polynucleotides or cloned into the same vector is introduced into the host cells already containing the daughter nucleic acids. Recombination is allowed to occur in the cells and the next generation of recombinants, or granddaughters are selected by the methods described above.

This cycle can be repeated a number of times until the nucleic acid or peptide having the desired characteristics is obtained. It is contemplated that in subsequent cycles, the population of mutated sequences which are added to the preferred hybrids may come from the parental hybrids or any subsequent generation.

In an alternative embodiment, the invention provides a method of conducting a "molecular" backcross of the obtained recombinant specific nucleic acid in order to eliminate any neutral mutations. Neutral mutations are those mutations which do not confer onto the nucleic acid or peptide the desired properties. Such mutations may however confer on the nucleic acid or peptide undesirable characteristics. Accordingly, it is desirable to eliminate such neutral mutations. The method of this invention provide a means of doing so.

In this embodiment, after the hybrid nucleic acid, having the desired characteristics, is obtained by the methods of the embodiments, the nucleic acid, the vector having the nucleic acid or the host cell containing the vector and nucleic acid is isolated.

The nucleic acid or vector is then introduced into the host cell with a large excess of the wild-type nucleic acid. The nucleic acid of the hybrid and the nucleic acid of the wild-type sequence are allowed to recombine. The resulting recombinants are placed under the same selection as the hybrid nucleic acid. Only those recombinants which retained the desired characteristics will be selected. Any silent mutations which do not provide the desired characteristics will be lost through recombination with the wild-type DNA. This cycle can be repeated a number of times until all of the silent mutations are eliminated.

Thus the methods of this invention can be used in a molecular backcross to eliminate unnecessary or silent mutations.

3.2.2. EXONUCLEASE-MEDIATED REASSEMBLY

In a particular embodiment, this invention provides for a method for shuffling, assembling, reassembling, recombining, &/or concatenating at least two polynucleotides to form a progeny polynucleotide (e.g. a chimeric progeny polynucleotide that can be expressed to produce a polypeptide or a gene pathway). In a particular embodiment, a double stranded polynucleotide end (e.g. two single stranded sequences hybridized to each other as hybridization partners) is treated with an exonuclease to liberate nucleotides from one of the two strands, leaving the remaining strand free of its original partner so that, if desired, the remaining strand may be used to achieve hybridization to another partner.

In a particular aspect, a double stranded polynucleotide end (that may be part of - or connected to - a polynucleotide or a nonpolynucleotide sequence) is subjected to a source of exonuclease activity. Serviceable sources of exonuclease activity may be an enzyme with 3' exonuclease activity, an enzyme with 5' exonuclease activity, an enzyme with both 3' exonuclease activity and 5' exonuclease activity, and any combination thereof. An exonuclease can be used to liberate nucleotides from one or both ends of a linear double stranded polynucleotide, and from one to all ends of a branched polynucleotide having more than two ends. The mechanism of action of this liberation is believed to be comprised of an enzymatically-catalyzed hydrolysis of terminal nucleotides, and can be allowed to proceed in a time-dependent fashion, allowing experimental control of the progression of the enzymatic process.

By contrast, a non-enzymatic step may be used to shuffle, assemble, reassemble, recombine, and/or concatenate polynucleotide building blocks that is comprised of subjecting a working sample to denaturing (or "melting") conditions (for example, by changing temperature, pH, and /or salinity conditions) so as to melt a working set of double stranded polynucleotides into single polynucleotide strands. For shuffling, it is desirable that the single polynucleotide strands participate to some extent in annealment with different hybridization partners (i.e. and not merely revert to exclusive reannealment between what were former partners before the denaturation step). The presence of the former hybridization partners in the reaction vessel, however, does not preclude, and may sometimes even favor, reannealment of a single stranded polynucleotide with its former partner, to recreate an original double stranded polynucleotide.

In contrast to this non-enzymatic shuffling step comprised of subjecting double stranded polynucleotide building blocks to denaturation, followed by annealment, the instant invention further provides an exonuclease-based approach requiring no denaturation – rather, the avoidance of denaturing conditions and the maintenance of double stranded polynucleotide substrates in annealed (i.e. non-denatured) state are necessary conditions for the action of exonucleases (e.g., exonuclease III and red alpha gene product). Additionally in contrast, the generation of single stranded polynucleotide sequences capable of hybridizing to other single stranded polynucleotide sequences is the result of covalent cleavage – and hence sequence destruction - in one of the hybridization partners. For example, an exonuclease III enzyme may be used to enzymatically liberate 3' terminal nucleotides in one hybridization strand (to achieve covalent hydrolysis in that polynucleotide strand); and this favors hybridization of the remaining single strand to a new partner (since its former partner was subjected to covalent cleavage).

By way of further illustration, a specific exonuclease, namely exonuclease III is provided herein as an example of a 3' exonuclease; however, other exonucleases may also be used, including enzymes with 5' exonuclease activity and enzymes with 3' exonuclease activity, and including enzymes not yet discovered and enzymes not yet developed. It is particularly appreciated that enzymes can be discovered, optimized (e.g. engineered by directed evolution), or both discovered and optimized specifically for the instantly disclosed approach that have more optimal rates &/or more highly specific activities &/or greater lack of unwanted activities. In fact it is expected that the instant invention may encourage the discovery &/or development of such designer enzymes. In sum, this invention may be practiced with a variety of currently available exonuclease enzymes, as well as enzymes not yet discovered and enzymes not yet developed.

The exonuclease action of exonuclease III requires a working double stranded polynucleotide end that is either blunt or has a 5' overhang, and the exonuclease action is comprised of enzymatically liberating 3' terminal nucleotides, leaving a single stranded 5' end that becomes longer and longer as the exonuclease action proceeds (see Figure 1). Any 5' overhangs produced by this approach may be used to hybridize to another single stranded polynucleotide sequence (which may also be a single stranded polynucleotide or a terminal overhang of a partially double stranded polynucleotide) that shares enough homology to allow hybridization. The ability of these exonuclease III-generated single stranded sequences (e.g. in 5' overhangs) to hybridize to other single stranded sequences allows two or more polynucleotides to be shuffled, assembled, reassembled, &/or concatenated.

Furthermore, it is appreciated that one can protect the end of a double stranded polynucleotide or render it susceptible to a desired enzymatic action of a serviceable exonuclease as necessary. For example, a double stranded polynucleotide end having a 3' overhang is not susceptible to the exonuclease action of exonuclease III. However, it may be rendered susceptible to the exonuclease action of exonuclease III by a variety of means; for example, it may be blunted by treatment with a polymerase, cleaved to provide a blunt end or a 5' overhang, joined (ligated or hybridized) to another double stranded polynucleotide to provide a blunt end or a 5' overhang, hybridized to a single stranded polynucleotide to provide a blunt end or a 5' overhang, or modified by any of a variety of means).

According to one aspect, an exonuclease may be allowed to act on one or on both ends of a linear double stranded polynucleotide and proceed to completion, to near completion, or to partial completion. When the exonuclease action is allowed to go to completion, the result will be that the length of each 5' overhang will extend far towards the middle region of the polynucleotide in the direction of what might be considered a "rendezvous point" (which may be somewhere near the polynucleotide midpoint). Ultimately, this results in the production of single stranded polynucleotides (that can become dissociated) that are each about half the length of the original double stranded polynucleotide (see Figure 1). Alternatively, an exonuclease-mediated reaction can be terminated before proceeding to completion.

Thus this exonuclease-mediated approach is serviceable for shuffling, assembling &/or reassembling, recombining, and concatenating polynucleotide building blocks, which polynucleotide building blocks can be up to ten bases long or tens of bases long or hundreds of bases long or thousands of bases long or tens of thousands of bases long or hundreds of thousands of bases long or millions of bases long or even longer.

This exonuclease-mediated approach is based on the action of double stranded DNA specific exodeoxyribonuclease activity of *E. coli* exonuclease III. Substrates for exonuclease III may be generated by subjecting a double stranded polynucleotide to fragmentation. Fragmentation may be achieved by mechanical means (e.g., shearing, sonication, etc.), by enzymatic means (e.g. using restriction enzymes), and by any combination thereof. Fragments of a larger polynucleotide may also be generated by polymerase-mediated synthesis.

Exonuclease III is a 28K monomeric enzyme, product of the *xth*A gene of *E. coli* with four known activities: exodeoxyribonuclease (alternatively referred to as exonuclease herein), RNaseH, DNA-3'-phosphatase, and AP endonuclease. The exodeoxyribonuclease activity is specific for double stranded DNA. The mechanism of action is thought to involve enzymatic hydrolysis of DNA from a 3' end progressively towards a 5' direction, with formation of nucleoside 5'-phosphates and a residual single strand. The enzyme does not display efficient hydrolysis of single stranded DNA, single-stranded RNA, or double-stranded RNA; however it degrades

RNA in an DNA-RNA hybrid releasing nucleoside 5'-phosphates. The enzyme also releases inorganic phosphate specifically from 3'phosphomonoester groups on DNA, but not from RNA or short oligonucleotides. Removal of these groups converts the terminus into a primer for DNA polymerase action.

Additional examples of enzymes with exonuclease activity include red-alpha and venom phosphodiesterases. Red alpha (*red* gene product (also referred to as lambda exonuclease) is of bacteriophage origin. The*red* gene is transcribed from the leftward promoter and its product is involved (24 kD) in recombination. Red alpha gene product acts processively from 5'-phosphorylated termini to liberate mononucleotides from duplex DNA (**Takahashi & Kobayashi**, 1990). Venom phosphodiesterases (Laskowski, 1980) is capable of rapidly opening supercoiled DNA.

3.2.3. NON-STOCHASTIC LIGATION REASSEMBLY

In one aspect, the present invention provides a non-stochastic method termed synthetic ligation reassembly (SLR), that is somewhat related to stochastic shuffling, save that the nucleic acid building blocks are not shuffled or concatenated or chimerized randomly, but rather are assembled non-stochastically.

A particularly glaring difference is that the instant SLR method does not depend on the presence of a high level of homology between polynucleotides to be shuffled. In contrast, prior methods, particularly prior stochastic shuffling methods require that presence of a high level of homology, particularly at coupling sites, between polynucleotides to be shuffled. Accordingly these prior methods favor the regeneration of the original progenitor molecules, and are suboptimal for generating large numbers of novel progeny chimeras, particularly full-length progenies. The instant invention, on the other hand, can be used to non-stochastically generate libraries (or sets) of progeny molecules comprised of over 10^{100} different chimeras. Conceivably, SLR can even be used to generate libraries comprised of over 10^{1000} different progeny chimeras with (no upper limit in sight).

Thus, in one aspect, the present invention provides a method, which method is non-stochastic, of producing a set of finalized chimeric nucleic acid molecules having an overall assembly order that is chosen by design, which method is comprised of the steps of generating by design a plurality of specific nucleic acid building blocks having serviceable mutually compatible ligatable ends, and assembling these nucleic acid building blocks, such that a designed overall assembly order is achieved.

The mutually compatible ligatable ends of the nucleic acid building blocks to be assembled are considered to be "serviceable" for this type of ordered assembly if they enable the building blocks to be coupled in predetermined orders. Thus, in one aspect, the overall assembly order in which the nucleic acid building blocks can be coupled is specified by the design of the ligatable ends and, if more than one assembly step is to be used, then the overall assembly order in which the nucleic acid building blocks can be coupled is also specified by the sequential order of the assembly step(s). Figure 4, Panel C illustrates an exemplary assembly process comprised of 2 sequential steps to achieve a designed (non-stochastic) overall assembly order for five nucleic acid building blocks. In a preferred embodiment of this invention, the annealed building pieces are treated with an enzyme, such as a ligase (e.g. T4 DNA ligase), achieve covalent bonding of the building pieces.

In a preferred embodiment, the design of nucleic acid building blocks is obtained upon analysis of the sequences of a set of progenitor nucleic acid templates that serve as a basis for producing a progeny set of finalized chimeric nucleic acid molecules. These progenitor nucleic acid templates thus serve as a source of sequence information that aids in the design of the nucleic acid building blocks that are to be mutagenized, i.e. chimerized or shuffled.

In one exemplification, this invention provides for the chimerization of a family of related genes and their encoded family of related products. In a particular exemplification, the encoded products are enzymes. As a representative list of families of enzymes which may be mutagenized in accordance with the aspects of the present invention, there may be mentioned, the following enzymes and their functions:

1 **Lipase/Esterase**

a. Enantioselective hydrolysis of esters (lipids)/ thioesters

1) Resolution of racemic mixtures

2) Synthesis of optically active acids or alcohols from *meso*-diesters

b. Selective syntheses

1) Regiospecific hydrolysis of carbohydrate esters

2) Selective hydrolysis of cyclic secondary alcohols

c. Synthesis of optically active esters, lactones, acids, alcohols

1) Transesterification of activated/nonactivated esters

2) Interesterification

3) Optically active lactones from hydroxyesters

4) Regio- and enantioselective ring opening of anhydrides

d. Detergents

e. Fat/Oil conversion

f. Cheese ripening

2 **Protease**

a. Ester/amide synthesis

b. Peptide synthesis

c. Resolution of racemic mixtures of amino acid esters

d. Synthesis of non-natural amino acids

e. Detergents/protein hydrolysis

3 **Glycosidase/Glycosyl transferase**

a. Sugar/polymer synthesis

b. Cleavage of glycosidic linkages to form mono, di-and oligosaccharides

c. Synthesis of complex oligosaccharides

d. Glycoside synthesis using UDP-galactosyl transferase

e. Transglycosylation of disaccharides, glycosyl fluorides, aryl galactosides

f. Glycosyl transfer in oligosaccharide synthesis

g. Diastereoselective cleavage of -glucosylsulfoxides

h. Asymmetric glycosylations

i. Food processing

j. Paper processing

4 **Phosphatase/Kinase**

a. Synthesis/hydrolysis of phosphate esters

1) Regio-, enantioselective phosphorylation

2) Introduction of phosphate esters

3) Synthesize phospholipid precursors

4) Controlled polynucleotide synthesis

b. Activate biological molecule

c. Selective phosphate bond formation without protecting groups

5 **Mono/Dioxygenase**

a. Direct oxyfunctionalization of unactivated organic substrates

b. Hydroxylation of alkane, aromatics, steroids

c. Epoxidation of alkenes

d. Enantioselective sulphoxidation

e. Regio- and stereoselective Bayer-Villiger oxidations

6 **Haloperoxidase**

a. Oxidative addition of halide ion to nucleophilic sites

b. Addition of hypohalous acids to olefinic bonds

c. Ring cleavage of cyclopropanes

d. Activated aromatic substrates converted to *ortho* and *para* derivatives

e. 1.3 diketones converted to 2-halo-derivatives

f. Heteroatom oxidation of sulfur and nitrogen containing substrates

g. Oxidation of enol acetates, alkynes and activated aromatic rings

7 **Lignin peroxidase/Diarylpropane peroxidase**

a. Oxidative cleavage of C-C bonds

b. Oxidation of benzylic alcohols to aldehydes

c. Hydroxylation of benzylic carbons

d. Phenol dimerization

e. Hydroxylation of double bonds to form diols

f. Cleavage of lignin aldehydes

8 **Epoxide hydrolase**

a. Synthesis of enantiomerically pure bioactive compounds
b. Regio- and enantioselective hydrolysis of epoxide
c. Aromatic and olefinic epoxidation by monooxygenases to form epoxides
d. Resolution of racemic epoxides
e. Hydrolysis of steroid epoxides

9 **Nitrile hydratase/nitrilase**

a. Hydrolysis of aliphatic nitriles to carboxamides
b. Hydrolysis of aromatic, heterocyclic, unsaturated aliphatic nitriles to corresponding acids
c. Hydrolysis of acrylonitrile
d. Production of aromatic and carboxamides, carboxylic acids (nicotinamide, picolinamide, isonicotinamide)
e. Regioselective hydrolysis of acrylic dinitrile
f. -amino acids from -hydroxynitriles

10 **Transaminase**

a. Transfer of amino groups into oxo-acids

11 **Amidase/Acylase**

a. Hydrolysis of amides, amidines, and other C-N bonds
b. Non-natural amino acid resolution and synthesis

These exemplifications, while illustrating certain specific aspects of the invention, do not portray the limitations or circumscribe the scope of the disclosed invention.

Thus according to one aspect of this invention, the sequences of a plurality of progenitor nucleic acid templates are aligned in order to select one or more demarcation points, which demarcation points can be located at an area of homology, and are comprised of one or more nucleotides, and which demarcation points are shared by at least two of the progenitor templates. The demarcation points can be used to delineate the boundaries of nucleic acid building blocks to be generated. Thus, the demarcation points identified and selected in the progenitor molecules serve as potential chimerization points in the assembly of the progeny molecules.

Preferably a serviceable demarcation point is an area of homology (comprised of at least one homologous nucleotide base) shared by at least two progenitor templates. More preferably a serviceable demarcation point is an area of homology that is shared by at least half of the progenitor templates. More preferably still a serviceable demarcation point is an area of homology that is shared by at least two thirds of the progenitor templates. Even more preferably a serviceable demarcation points is an area of homology that is shared by at least three fourths of the progenitor templates. Even more preferably still a serviceable demarcation points is an area of homology that is shared by at almost all of the progenitor templates. Even more preferably still a serviceable demarcation point is an area of homology that is shared by all of the progenitor templates.

The process of designing nucleic acid building blocks and of designing the mutually compatible ligatable ends of the nucleic acid building blocks to be assembled is illustrated in Figures 6 and 7. As shown, the alignment of a set of progenitor templates reveals several naturally occurring demarcation points, and the identification of demarcation points shared by these templates helps to non-stochastically determine the building blocks to be generated and used for the generation of the progeny chimeric molecules.

In a preferred embodiment, this invention provides that the ligation reassembly process is performed exhaustively in order to generate an exhaustive library. In other words, all possible ordered combinations of the nucleic acid building blocks are represented in the set of finalized chimeric nucleic acid molecules. At the same time, in a particularly preferred embodiment, the assembly order (i.e. the order of assembly of each building block in the 5' to 3' sequence of each finalized chimeric nucleic acid) in each combination is by design (or non-stochastic). Because of the non-stochastic nature of this invention, the possibility of unwanted side products is greatly reduced.

In another preferred embodiment, this invention provides that the ligation reassembly process is performed systematically, for example in order to generate a systematically compartmentalized library, with compartments that can be screened systematically, e.g. one by one. In other words this invention provides that, through the selective and judicious use of specific nucleic acid building blocks, coupled with the selective and judicious use of sequentially stepped assembly reactions, an experimental design can be achieved where specific sets of progeny products are made in each of several reaction vessels. This allows a systematic examination and screening procedure to be performed. Thus, it allows a potentially very large number of progeny molecules to be examined systematically in smaller groups.

Because of its ability to perform chimerizations in a manner that is highly flexible yet exhaustive and systematic as well, particularly when there is a low level of homology among the progenitor molecules, the instant invention provides for the generation of a library (or set) comprised of a large number of progeny molecules. Because of the non-stochastic nature of the instant ligation reassembly invention, the progeny molecules generated preferably comprise a library of finalized chimeric nucleic acid molecules having an overall assembly order that is chosen by design. In a particularly preferred embodiment of this invention, such a generated library is comprised of preferably greater than 10^3 different progeny molecular species, more preferably greater than 10^5 different progeny molecular species, more preferably still greater than 10^{10} different progeny molecular species, more preferably still greater than 10^{15} different progeny molecular species, more preferably still greater than 10^{20} different progeny molecular species, more preferably still greater than 10^{30} different progeny molecular species, more preferably still greater than 10^{40} different progeny molecular species, more preferably still greater than 10^{50} different progeny molecular species, more preferably still greater than 10^{60} different progeny molecular species, more preferably still greater than 10^{70} different progeny molecular species, more preferably still greater than 10^{80} different progeny molecular species, more preferably still greater than 10^{100} different progeny molecular species, more preferably still greater than 10^{110} different progeny molecular species, more preferably still greater than 10^{120} different progeny molecular species, more preferably still greater than 10^{130} different progeny molecular species, more preferably still greater than 10^{140} different progeny molecular species, more preferably still greater than 10^{150} different progeny molecular species, more preferably still greater than 10^{175} different progeny molecular species, more preferably still greater than 10^{200} different progeny molecular species, more preferably still greater than 10^{300} different progeny molecular species, more preferably still greater than 10^{400} different progeny molecular species, more preferably still greater than 10^{500} different progeny molecular species, and even more preferably still greater than 10^{1000} different progeny molecular species.

In one aspect, a set of finalized chimeric nucleic acid molecules, produced as described is comprised of a polynucleotide encoding a polypeptide. According to one preferred embodiment, this polynucleotide is a gene, which may be a man-made gene. According to another preferred embodiment, this polynucleotide is a gene pathway, which may be a man-made gene pathway. This invention provides that one or more man-made genes generated by this invention may be incorporated into a man-made gene pathway, such as a pathway operable in a eukaryotic organism (including a plant).

It is appreciated that the power of this invention is exceptional, as there is much freedom of choice and control regarding the selection of demarcation points, the size and number of the nucleic acid building blocks, and the size and design of the couplings. It is appreciated, furthermore, that the requirement for intermolecular homology is highly relaxed for the operability of this invention. In fact, demarcation points can even be chosen in areas of little or no intermolecular homology. For example, because of codon wobble, i.e. the degeneracy of codons, nucleotide substitutions can be introduced into nucleic acid building blocks without altering the amino acid originally encoded in the corresponding progenitor template. Alternatively, a codon can be altered such that the coding for an originally amino acid is altered. This invention provides that such substitutions can be introduced into the nucleic acid building block in order to increase the incidence of intermolecularly homologous demarcation points and thus to allow an increased number of couplings to be achieved among the building blocks, which in turn allows a greater number of progeny chimeric molecules to be generated.

In another exemplifaction, the synthetic nature of the step in which the building blocks are generated allows the design and introduction of nucleotides (e.g. one or more nucleotides, which may be, for example, codons or introns or regulatory sequences) that can later be optionally removed in an in vitro process (e.g. by mutageneis) or in an in vivo process (e.g. by utilizing the gene splicing ability of a host organism). It is appreciated that in many instances the introduction of these nucleotides may also be desirable for many other reasons in addition to the potential benefit of creating a serviceable demarcation point.

Thus, according to another embodiment, this invention provides that a nucleic acid building block can be used to introduce an intron. Thus, this invention provides that functional introns may be introduced into a man-made gene of this invention. This invention also provides that functional introns may be introduced into a man-made gene pathway of this invention. Accordingly, this invention provides for the generation of a chimeric polynucleotide that is a man-made gene containing one (or more) artificially introduced intron(s).

Accordingly, this invention also provides for the generation of a chimeric polynucleotide that is a man-made gene pathway containing one (or more) artificially introduced intron(s). Preferably, the artificially introduced intron(s) are functional in one or more host cells for gene splicing much in the way that naturally-occurring introns serve functionally in gene splicing. This invention provides a process of producing man-made intron-containing polynucleotides to be introduced into host organisms for recombination and/or splicing.

The ability to achieve chimerizations, using couplings as described herein, in areas of little or no homology among the progenitor molecules, is particularly useful, and in fact critical, for the assembly of novel gene pathways. This invention thus provides for the generation of novel man-made gene pathways using synthetic ligation reassembly. In a particular aspect, this is achieved by the introduction of regulatory sequences, such as promoters, that are operable in an intended host, to confer operability to a novel gene pathway when it is introduced into the intended host. In a particular exemplification, this invention provides for the generation of novel man-made gene pathways that is operable in a plurality of intended hosts (e.g. in a microbial organism as well as in a plant cell). This can be achieved, for example, by the introduction of a plurality of regulatory sequences, comprised of a regulatory sequence that is operable in a first intended host and a regulatory sequence that is operable in a second intended host. A similar process can be performed to achieve operability of a gene pathway in a third intended host species, etc. The number of intended host species can be each integer from 1 to 10 or alternatively over 10. Alternatively, for example, operability of a gene pathway in a plurality of intended hosts can be achieved by the introduction of a regulatory sequence having intrinsic operability in a plurality of intended hosts.

Thus, according to a particular embodiment, this invention provides that a nucleic acid building block can be used to introduce a regulatory sequence, particularly a regulatory sequence for gene expression. Preferred regulatory sequences include, but are not limited to, those that are man-made, and those found in archeal, bacterial, eukaryotic (including mitochondrial), viral, and prionic or prion-like organisms. Preferred regulatory sequences include but are not limited to, promoters, operators, and activator binding sites. Thus, this invention provides that functional regulatory sequences may be introduced into a man-made gene of this invention. This invention also provides that functional regulatory sequences may be introduced into a man-made gene pathway of this invention.

Accordingly, this invention provides for the generation of a chimeric polynucleotide that is a man-made gene containing one (or more) artificially introduced regulatory sequence(s). Accordingly, this invention also provides for the generation of a chimeric polynucleotide that is a man-made gene pathway containing one (or more) artificially introduced regulatory sequence(s). Preferably, an artificially introduced regulatory sequence(s) is operatively linked to one or more genes in the man-made polynucleotide, and are functional in one or more host cells.

Preferred bacterial promoters that are serviceable for this invention include lacI, lacZ, T3, T7, gpt, lambda P_R, P_L and trp. Serviceable eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Particular plant regulatory sequences include promoters active in directing transcription in plants, either constitutively or stage and/or tissue specific, depending on the use of the plant or parts thereof. These promoters include, but are not limited to promoters showing constitutive expression, such as the 35S promoter of Cauliflower Mosaic Virus (CaMV) (**Guilley** et al., 1982), those for leaf-specific expression, such as the promoter of the ribulose bisphosphate carboxylase small subunit gene (**Coruzzi** et al., 1984), those for root-specific expression, such as the promoter from the glutamin synthase gene (**Tingey** et al., 1987), those for seed-specific expression, such as the cruciferin A promoter from *Brassica napus* (**Ryan** et al., 1989), those for tuber-specific expression, such as the class-I patatin promoter from potato (**Rocha-Sasa** et al., 1989; **Wenzler** et al., 1989) or those for fruit-specific expression, such as the polygalacturonase (PG) promoter from tomato (**Bird** et al., 1988).

Other regulatory sequences that are preferred for this invention include terminator sequences and polyadenylation signals and any such sequence functioning as such in plants, the choice of which is within the level of the skilled artisan. An example of such sequences is the 3′ flanking region of the nopaline synthase (nos) gene of *Agrobacterium tumefaciens* (**Bevan**, 1984). The regulatory sequences may also include enhancer sequences, such as found in the 35S promoter of CaMV, and mRNA stabilizing sequences such as the leader sequence of Alfalfa Mosaic Cirus (AlMV) RNA4 (**Brederode** et al., 1980) or any other sequences functioning in a like manner.

Man-made genes produced using this invention can also serve as a substrate for recombination with another nucleic acid. Likewise, a man-made gene pathway produced using this invention can also serve as a substrate for recombination with another nucleic acid. In a preferred instance, the recombination is facilitated by, or occurs at, areas of homology between the man-made intron-containing gene and a nucleic acid with serves as a recombination partner. In a particularly preferred instance, the recombination partner may also be a nucleic acid generated by this invention, including a man-made gene or a man-made gene pathway. Recombination may be facilitated by or may occur at areas of homology that exist at the one (or more) artificially introduced intron(s) in the man-made gene.

The synthetic ligation reassembly method of this invention utilizes a plurality of nucleic acid building blocks, each of which preferably has two ligatable ends. The two ligatable ends on each nucleic acid building block may be two blunt ends (i.e. each having an overhang of zero nucleotides), or preferably one blunt end and one overhang, or more preferably still two overhangs.

A serviceable overhang for this purpose may be a 3' overhang or a 5' overhang. Thus, a nucleic acid building block may have a 3' overhang or alternatively a 5' overhang or alternatively two 3' overhangs or alternatively two 5' overhangs. The overall order in which the nucleic acid building blocks are assembled to form a finalized chimeric nucleic acid molecule is determined by purposeful experimental design and is not random.

According to one preferred embodiment, a nucleic acid building block is generated by chemical synthesis of two single-stranded nucleic acids (also referred to as single-stranded oligos) and contacting them so as to allow them to anneal to form a double-stranded nucleic acid building block.

A double-stranded nucleic acid building block can be of variable size. The sizes of these building blocks can be small or large depending on the choice of the experimenter. Preferred sizes for building block range from 1 base pair (not including any overhangs) to 100,000 base pairs (not including any overhangs). Other preferred size ranges are also provided, which have lower limits of from 1 bp to 10,000 bp (including every integer value in between), and upper limits of from 2 bp to 100, 000 bp (including every integer value in between).

It is appreciated that current methods of polymerase-based amplification can be used to generate double-stranded nucleic acids of up to thousands of base pairs, if not tens of thousands of base pairs, in length with high fidelity. Chemical synthesis (e.g. phosphoramidite-based) can be used to generate nucleic acids of up to hundreds of nucleotides in length with high fidelity; however, these can be assembled, e.g. using overhangs or sticky ends, to form double-stranded nucleic acids of up to thousands of base pairs, if not tens of thousands of base pairs, in length if so desired.

A combination of methods (e.g. phosphoramidite-based chemical synthesis and PCR) can also be used according to this invention. Thus, nucleic acid building block made by different methods can also be used in combination to generate a progeny molecule of this invention.

The use of chemical synthesis to generate nucleic acid building blocks is particularly preferred in this invention & is advantageous for other reasons as well, including procedural safety and ease. No cloning or harvesting or actual handling of any biological samples is required. The design of the nucleic acid building blocks can be accomplished on paper. Accordingly, this invention teaches an advance in procedural safety in recombinant technologies.

Nonetheless, according to one preferred embodiment, a double-stranded nucleic acid building block according to this invention may also be generated by polymerase-based amplification of a polynucleotide template. In a non-limiting exemplification, as illustrated in Figure 2, a first polymerase-based amplification reaction using a first set of primers, F_2 and R_1, is used to generate a blunt-ended product (labeled Reaction 1, Product 1), which is essentially identical to Product A. A second polymerase-based amplification reaction using a second set of primers, F_1 and R_2, is used to generate a blunt-ended product (labeled Reaction 2, Product 2), which is essentially identical to Product B. These two products are mixed and allowed to melt and anneal, generating potentially useful double-stranded nucleic acid building blocks with two overhangs. In the example of Fig. 2, the product with the 3' overhangs (Product C) is selected by nuclease-based degradation of the other 3 products using a 3' acting exonuclease, such as exonuclease III. It is appreciated that a 5' acting exonuclease (e.g. red alpha) may be also be used, for example to select Product D instead. It is also appreciated that other selection means can also be used, including hybridization-based means, and that these means can incorporate a further means, such as a magnetic bead-based means, to facilitate separation of the desired product.

Many other methods exist by which a double-stranded nucleic acid building block can be generated that is serviceable for this invention; and these are known in the art and can be readily performed by the skilled artisan.

According to particularly preferred embodiment, a double-stranded nucleic acid building block that is serviceable for this invention is generated by first generating two single stranded nucleic acids and allowing them to anneal to form a double-stranded nucleic acid building block. The two strands of a double-stranded nucleic acid building block may be complementary at every nucleotide apart from any that form an overhang; thus containing no mismatches, apart from any overhang(s). According to another embodiment, the two strands of a double-stranded nucleic acid building block are complementary at fewer than every nucleotide apart from any that form an overhang. Thus, according to this embodiment, a double-stranded nucleic acid building block can be used to introduce codon degeneracy. Preferably the codon degeneracy is introduced using the site-saturation mutagenesis described herein, using one or more N,N,G/T cassettes or alternatively using one or more N,N,N cassettes.

Contained within an exemplary experimental design for achieving an ordered assembly according to this invention are:

1) The design of specific nucleic acid building blocks.
2) The design of specific ligatable ends on each nucleic acid building block.
3) The design of a particular order of assembly of the nucleic acid building blocks.

An overhang may be a 3' overhang or a 5' overhang. An overhang may also have a terminal phosphate group or alternatively may be devoid of a terminal phosphate group (having, e.g., a hydroxyl group instead). An overhang may be comprised of any number of nucleotides. Preferably an overhang is comprised of 0 nucleotides (as in a blunt end) to 10,000 nucleotides. Thus, a wide range of overhang sizes may be serviceable. Accordingly, the lower limit may be each integer from 1-200 and the upper limit may be each integer from 2-10,000. According to a particular exemplification, an overhang may consist of anywhere from 1 nucleotide to 200 nucleotides (including every integer value in between).

The final chimeric nucleic acid molecule may be generated by sequentially assembling 2 or more building blocks at a time until all the designated building blocks have been assembled. A working sample may optionally be subjected to a process for size selection or purification or other selection or enrichment process between the performance of two assembly steps. Alternatively, the final chimeric nucleic acid molecule may be generated by assembling all the designated building blocks at once in one step.

Utility

The *in vivo* recombination method of this invention can be performed blindly on a pool of unknown hybrids or alleles of a specific polynucleotide or sequence. However, it is not necessary to know the actual DNA or RNA sequence of the specific polynucleotide.

The approach of using recombination within a mixed population of genes can be useful for the generation of any useful proteins, for example, interleukin I, antibodies, tPA and growth hormone. This approach may be used to generate proteins having altered specificity or activity. The approach may also be useful for the generation of hybrid nucleic acid sequences, for example, promoter regions, introns, exons, enhancer sequences, 3' untranslated regions or 5' untranslated regions of genes. Thus this approach may be used to generate genes having increased rates of expression. This approach may also be useful in the study of repetitive DNA sequences. Finally, this approach may be useful to mutate ribozymes or aptamers.

Scaffold-like regions separating regions of diversity in proteins may be particularly suitable for the methods of this invention. The conserved scaffold determines the overall folding by self-association, while displaying relatively unrestricted loops that mediate the specific binding. Examples of such scaffolds are the immunoglobulin beta barrel, and the four-helix bundle. The methods of this invention can be used to create scaffold-like proteins with various combinations of mutated sequences for binding.

The equivalents of some standard genetic matings may also be performed by the methods of this invention. For example, a "molecular" backcross can be performed by repeated mixing of the hybrid's nucleic acid with the wild-type nucleic acid while selecting for the mutations of interest. As in traditional breeding, this approach can be used to combine phenotypes from different sources into a background of choice. It is useful, for example, for the removal of neutral mutations that affect unselected characteristics (i.e. immunogenicity). Thus it can be useful to determine which mutations in a protein are involved in the enhanced biological activity and which are not.

3.2.4. END-SELECTION

This invention provides a method for selecting a subset of polynucleotides from a starting set of polynucleotides, which method is based on the ability to discriminate one or more selectable features (or selection markers) present anywhere in a working polynucleotide, so as to allow one to perform selection for (positive selection) &/or against (negative selection) each selectable polynucleotide. In a preferred aspect, a method is provided termed end-selection, which method is based on the use of a selection marker located in part or entirely in a terminal region of a selectable polynucleotide, and such a selection marker may be termed an "end-selection marker".

End-selection may be based on detection of naturally occurring sequences or on detection of sequences introduced experimentally (including by any mutagenesis procedure mentioned herein and not mentioned herein) or on both, even within the same polynucleotide. An end-selection marker can be a structural selection marker or a functional selection marker or both a structural and a functional selection marker. An end-selection marker may be comprised of a polynucleotide sequence or of a polypeptide sequence or of any chemical structure or of any biological or biochemical tag, including markers that can be selected using methods based on the detection of radioactivity, of enzymatic activity, of fluorescence, of any optical feature, of a magnetic property (e.g. using magnetic beads), of immunoreactivity, and of hybridization.

End-selection may be applied in combination with any method serviceable for performing mutagenesis. Such mutagenesis methods include, but are not limited to, methods described herein (supra and infra). Such methods include, by way of non-limiting exemplification, any method that may be referred herein or by others in the art by any of the following terms: "saturation mutagenesis", "shuffling", "recombination", "re-assembly", "error-prone PCR", "assembly PCR", "sexual PCR", "crossover PCR", "oligonucleotide primer-directed mutagenesis", "recursive (&/or exponential) ensemble mutagenesis (see **Arkin** and **Youvan**, 1992)", "cassette mutagenesis", "in vivo mutagenesis", and "in vitro mutagenesis". Moreover, end-selection may be performed on molecules produced by any mutagenesis &/or amplification method (see, e.g., **Arnold**, 1993; **Caldwell** and **Joyce**, 1992; **Stemmer**, 1994; following which method it is desirable to select for (including to screen for the presence of) desirable progeny molecules.

In addition, end-selection may be applied to a polynucleotide apart from any mutagenesis method. In a preferred embodiment, end-selection, as provided herein, can be used in order to facilitate a cloning step, such as a step of ligation to another polynucleotide (including ligation to a vector). This invention thus provides for end-selection as a serviceable means to facilitate library construction, selection &/or enrichment for desirable polynucleotides, and cloning in general.

In a particularly preferred embodiment, end-selection can be based on (positive) selection for a polynucleotide; alternatively end-selection can be based on (negative) selection against a polynucleotide; and alternatively still, end-selection can be based on both (positive) selection for, and on (negative) selection against, a polynucleotide. End-selection, along with other methods of selection &/or screening, can be performed in an iterative fashion, with any combination of like or unlike selection &/or screening methods and serviceable mutagenesis methods, all of which can be performed in an iterative fashion and in any order, combination, and permutation.

It is also appreciated that, according to one embodiment of this invention, end-selection may also be used to select a polynucleotide that is at least in part: circular (e.g. a plasmid or any other circular vector or any other polynucleotide that is partly circular), &/or branched, &/or modified or substituted with any chemical group or moiety. In accord with this embodiment, a polynucleotide may be a circular molecule comprised of an intermediate or central region, which region is flanked on a 5' side by a 5' flanking region (which, for the purpose of end-selection, serves in like manner to a 5' terminal region of a non-circular polynucleotide) and on a 3' side by a 3' terminal region (which, for the purpose of end-selection, serves in like manner to a 3' terminal region of a non-circular polynucleotide). As used in this non-limiting exemplification, there may be sequence overlap between any two regions or even among all three regions.

In one non-limiting aspect of this invention, end-selection of a linear polynucleotide is performed using a general approach based on the presence of at least one end-selection marker located at or near a polynucleotide end or terminus (that can be either a 5' end or a 3' end). In one particular non-limiting exemplification, end-selection is based on selection for a specific sequence at or near a terminus such as, but not limited to, a sequence recognized by an enzyme that recognizes a polynucleotide sequence. An enzyme that recognizes and catalyzes a chemical modification of a polynucleotide is referred to herein as a polynucleotide-acting enzyme. In a preferred embodiment, serviceable polynucleotide-acting enzymes are exemplified non-exclusively by enzymes with polynucleotide-cleaving activity, enzymes with polynucleotide-methylating activity, enzymes with polynucleotide-ligating activity, and enzymes with a plurality of distinguishable enzymatic activities (including non-exclusively, e.g., both polynucleotide-cleaving activity and polynucleotide-ligating activity).

Relevant polynucleotide-acting enzymes thus also include any commercially available or non-commercially available polynucleotide endonucleases and their companion methylases including those catalogued at the website http://www.neb.com/rebase, and those mentioned in the following cited reference (**Roberts** and **Macelis**, 1996). Preferred polynucleotide endonucleases include – but are not limited to – type II restriction enzymes (including type IIS), and include enzymes that cleave both strands of a double stranded polynucleotide (e.g. *Not* I, which cleaves both strands at 5'...GC/GGCCGC...3') and enzymes that cleave only one strand of a double stranded polynucleotide, i.e. enzymes that have polynucleotide-nicking activity, (e.g. N. *Bst*NB I, which cleaves only one strand at 5'...GAGTCNNNN/N...3'). Relevant polynucleotide-acting enzymes also include type III restriction enzymes.

It is appreciated that relevant polynucleotide-acting enzymes also include any enzymes that may be developed in the future, though currently unavailable, that are serviceable for generating a ligation compatible end, preferably a sticky end, in a polynucleotide.

In one preferred exemplification, a serviceable selection marker is a restriction site in a polynucleotide that allows a corresponding type II (or type IIS) restriction enzyme to cleave an end of the polynucleotide so as to provide a ligatable end (including a blunt end or alternatively a sticky end with at least a one base overhang) that is serviceable for a desirable ligation reaction without cleaving the polynucleotide internally in a manner that destroys a desired internal sequence in the polynucleotide. Thus it is provided that, among relevant restriction sites, those sites that do not occur internally (i.e. that do not occur apart from the termini) in a specific working polynucleotide are preferred when the use of a corresponding restriction enzyme(s) is not intended to cut the working polynucleotide internally. This allows one to perform restriction digestion reactions to completion or to near completion without incurring unwanted internal cleavage in a working polynucleotide.

According to a preferred aspect, it is thus preferable to use restriction sites that are not contained, or alternatively that are not expected to be contained, or alternatively that are unlikely to be contained (e.g. when sequence information regarding a working polynucleotide is incomplete) internally in a polynucleotide to be subjected to end-selection. In accordance with this aspect, it is appreciated that restriction sites that occur relatively infrequently are usually preferred over those that occur more frequently. On the other hand it is also appreciated that there are occasions where internal cleavage of a polypeptide is desired, e.g. to achieve recombination or other mutagenic procedures along with end-selection.

In accord with this invention, it is also appreciated that methods (e.g. mutagenesis methods) can be used to remove unwanted internal restriction sites. It is also appreciated that a partial digestion reaction (i.e. a digestion reaction that proceeds to partial completion) can be used to achieve digestion at a recognition site in a terminal region while sparing a susceptible restriction site that occurs internally in a polynucleotide and that is recognized by the same enzyme. In one aspect, partial digest are useful because it is appreciated that certain enzymes show preferential cleavage of the same recognition sequence depending on the location and environment in which the recognition sequence occurs. For example, it is appreciated that, while lambda DNA has 5 *Eco*R I sites, cleavage of the site nearest to the right terminus has been reported to occur 10 times faster than the sites in the middle of the molecule. Also, for example, it has been reported that, while *Sac* II has four sites on lambda DNA, the three clustered centrally in lambda are cleaved 50 times faster than the remaining site near the terminus (at nucleotide 40,386). Summarily, site preferences have been reported for various enzymes by many investigators (e.g., **Thomas** and **Davis**, 1975; **Forsblum** et al, 1976; **Nath** and **Azzolina**, 1981; **Brown** and **Smith**, 1977; **Gingeras** and **Brooks**, 1983; **Krüger** et al, 1988; **Conrad** and **Topal**, 1989; **Oller** et al, 1991; **Topal**, 1991; and **Pein**, 1991; to name but a few). It is appreciated that any empirical observations as well as any mechanistic understandings of site preferences by any serviceable polynucleotide-acting enzymes, whether currently available or to be procured in the future, may be serviceable in end-selection according to this invention.

It is also appreciated that protection methods can be used to selectively protect specified restriction sites (e.g. internal sites) against unwanted digestion by enzymes that would otherwise cut a working polypeptide in response to the presence of those sites; and that such protection methods include modifications such as methylations and base substitutions (e.g. U instead of T) that inhibit an unwanted enzyme activity. It is appreciated that there are limited numbers of available restriction enzymes that are rare enough (e.g. having very long recognition sequences) to create large (e.g. megabase-long) restriction fragments, and that protection approaches (e.g. by methylation) are serviceable for increasing the rarity of enzyme cleavage sites. The use of M.*Fnu* II (mCGCG) to increase the apparent rarity of *Not* I approximately twofold is but one example among many (**Qiang** et al, 1990; **Nelson** et al, 1984; **Maxam** and **Gilbert**, 1980; **Raleigh** and **Wilson**, 1986).

According to a preferred aspect of this invention, it is provided that, in general, the use of rare restriction sites is preferred. It is appreciated that, in general, the frequency of occurrence of a restriction site is determined by the number of nucleotides contained therein, as well as by the ambiguity of the base requirements contained therein. Thus, in a non-limiting exemplification, it is appreciated that, in general, a restriction site composed of, for example, 8 specific nucleotides (e.g. the *Not* I site or GC/GGCCGC, with an estimated relative occurrence of 1 in 4^8, i.e. 1 in 65,536, random 8-mers) is relatively more infrequent than one composed of, for example, 6 nucleotides (e.g. the *Sma* I site or CCC/GGG, having an estimated relative occurrence of 1 in 4^6, i.e. 1 in 4,096, random 6-mers), which in turn is relatively more infrequent than one composed of, for example, 4 nucleotides (e.g. the *Msp* I site or C/CGG, having an estimated relative occurrence of 1 in 4^4, i.e. 1 in 256, random 4-mers). Moreover, in another non-limiting exemplification, it is appreciated that, in general, a restriction site having no ambiguous (but only specific) base requirements (e.g. the *Fin* I site or GTCCC, having an estimated relative occurrence of 1 in 4^5, i.e. 1 in 1024, random 5-mers) is relatively more infrequent than one having an ambiguous W (where W = A or T) base requirement (e.g. the *Ava* II site or G/GWCC, having an estimated relative occurrence of 1 in 4x4x2x4x4 - i.e. 1 in 512 – random 5-mers), which in turn is relatively more infrequent than one having an ambiguous N (where N = A or C or G or T) base requirement (e.g. the *Asu* I site or G/GNCC, having an estimated relative occurrence of 1 in 4x4x1x4x4, i.e. 1 in 256 – random 5-mers). These relative occurrences are considered general estimates for actual polynucleotides, because it is appreciated that specific nucleotide bases (not to mention specific nucleotide sequences) occur with dissimilar frequencies in specific polynucleotides, in specific species of organisms, and in specific groupings of organisms. For example, it is appreciated that the % G+C contents of different species of organisms are often very different and wide ranging.

The use of relatively more infrequent restriction sites as a selection marker include - in a non-limiting fashion - preferably those sites composed at least a 4 nucleotide sequence, more preferably those composed of at least a 5 nucleotide sequence, more preferably still those composed at least a 6 nucleotide sequence (e.g. the *Bam*H I site or G/GATCC, the *Bgl* II site or A/GATCT, the *Pst* I site or CTGCA/G, and the *Xba* I site or T/CTAGA), more preferably still those composed at least a 7 nucleotide sequence, more preferably still those composed of an 8 nucleotide sequence nucleotide sequence (e.g. the *Asc* I site or GG/CGCGCC, the *Not* I site or GC/GGCCGC, the *Pac* I site or TTAAT/TAA, the *Pme* I site or GTTT/AAAC, the *Srf* I site or GCCC/GGGC, the *Sse*838 I site or CCTGCA/GG, and the *Swa* I site or ATTT/AAAT), more preferably still those composed of a 9 nucleotide sequence, and even more preferably still those composed of at least a 10 nucleotide sequence (e.g. the *Bsp*G I site or CG/CGCTGGAC). It is further appreciated that some restriction sites (e.g. for class IIS enzymes) are comprised of a portion of relatively high specificity (i.e. a portion containing a principal determinant of the frequency of occurrence of the restriction site) and a portion of relatively low specificity; and that a site of cleavage may or may not be contained within a portion of relatively low specificity. For example, in the *Eco*57 I site or CTGAAG(16/14), there is a portion of relatively high specificity (i.e. the CTGAAG portion) and a portion of relatively low specificity (i.e. the N16 sequence) that contains a site of cleavage.

In another preferred embodiment of this invention, a serviceable end-selection marker is a terminal sequence that is recognized by a polynucleotide-acting enzyme that recognizes a specific polynucleotide sequence. In a preferred aspect of this invention, serviceable polynucleotide-acting enzymes also include other enzymes in addition to classic type II restriction enzymes. According to this preferred aspect of this invention, serviceable polynucleotide-acting enzymes also include gyrases, helicases, recombinases, relaxases, and any enzymes related thereto.

Among preferred examples are topoisomerases (which have been categorized by some as a subset of the gyrases) and any other enzymes that have polynucleotide-cleaving activity (including preferably polynucleotide-nicking activity) &/or polynucleotide-ligating activity. Among preferred topoisomerase enzymes are topoisomerase I enzymes, which is available from many commercial sources (Epicentre Technologies, Madison, WI; Invitrogen, Carlsbad, CA; Life Technologies, Gathesburg, MD) and conceivably even more private sources. It is appreciated that similar enzymes may be developed in the future that are serviceable for end-selection as provided herein. A particularly preferred topoisomerase I enzyme is a topoisomerase I enzyme of vaccinia virus origin, that has a specific recognition sequence (e.g. 5'...AAGGG...3') and has both polynucleotide-nicking activity and polynucleotide-ligating activity. Due to the specific nicking-activity of this enzyme (cleavage of one strand), internal recognition sites are not prone to polynucleotide destruction resulting from the nicking activity (but rather remain annealed) at a temperature that causes denaturation of a terminal site that has been nicked. Thus for use in end-selection, it is preferable that a nicking site for topoisomerase-based end-selection be no more than 100 nucleotides from a terminus, more preferably no more than 50 nucleotides from a terminus, more preferably still no more than 25 nucloetides from a terminus, even more preferably still no more than 20 nucleotides from a terminus, even more preferably still no more than 15 nucleotides from a terminus, even more preferably still no more than 10 nucleotides from a terminus, even more preferably still no more than 8 nucleotides from a terminus, even more preferably still no more than 6 nucleotides from a terminus, and even more preferably still no more than 4 nucleotides from a terminus.

In a particularly preferred exemplification that is non-limiting yet clearly illustrative, it is appreciated that when a nicking site for topoisomerase-based end-selection is 4 nucleotides from a terminus, nicking produces a single stranded oligo of 4 bases (in a terminal region) that can be denatured from its complementary strand in an end-selectable polynucleotide; this provides a sticky end (comprised of 4 bases) in a polynucleotide that is serviceable for an ensuing ligation reaction. To accomplish ligation to a cloning vector (preferably an expression vector), compatible sticky ends can be generated in a cloning vector by any means including by restriction enzyme-based means. The terminal nucleotides (comprised of 4 terminal bases in this specific example) in an end-selectable polynucleotide terminus are thus wisely chosen to provide compatibility with a sticky end generated in a cloning vector to which the polynucleotide is to be ligated.

On the other hand, internal nicking of an end-selectable polynucleotide, e.g. 500 bases from a terminus, produces a single stranded oligo of 500 bases that is not easily denatured from its complementary strand, but rather is serviceable for repair (e.g. by the same topoisomerase enzyme that produced the nick).

This invention thus provides a method - e.g. that is vaccinia topoisomerase-based &/or type II (or IIS) restriction endonuclease-based &/or type III restriction endonuclease-based &/or nicking enzyme-based (e.g. using N. *Bst*NB I) – for producing a sticky end in a working polynucleotide, which end is ligation compatible, and which end can be comprised of at least a 1 base overhang. Preferably such a sticky end is comprised of at least a 2-base overhang, more preferably such a sticky end is comprised of at least a 3-base overhang, more preferably still such a sticky end is comprised of at least a 4-base overhang, even more preferably still such a sticky end is comprised of at least a 5-base overhang, even more preferably still such a sticky end is comprised of at least a 6-base overhang. Such a sticky end may also be comprised of at least a 7-base overhang, or at least an 8-base overhang, or at least a 9-base overhang, or at least a 10-base overhang, or at least 15-base overhang, or at least a 20-base overhang, or at least a 25-base overhang, or at least a 30-base overhang. These overhangs can be comprised of any bases, including A, C, G, or T.

It is appreciated that sticky end overhangs introduced using topoisomerase or a nicking enzyme (e.g. using N. *Bst*NB I) can be designed to be unique in a ligation environment, so as to prevent unwanted fragment reassemblies, such as self-dimerizations and other unwanted concatamerizations.

According to one aspect of this invention, a plurality of sequences (which may but do not necessarily overlap) can be introduced into a terminal region of an end-selectable polynucleotide by the use of an oligo in a polymerase-based reaction. In a relevant, but by no means limiting example, such an oligo can be used to provide a preferred 5' terminal region that is serviceable for topoisomerase I-based end-selection, which oligo is comprised of: a 1-10 base sequence that is convertible into a sticky end (preferably by a vaccinia topoisomerase I), a ribosome binding site (i.e. and "RBS", that is preferably serviceable for expression cloning), and optional linker sequence followed by an ATG start site and a template-specific sequence of 0-100 bases (to facilitate annealment to the template in the polymerase-based reaction). Thus, according to this example, a serviceable oligo (which may be termed a forward primer) can have the sequence: 5'[terminal sequence = $(N)_{1-10}$][topoisomerase I site &

RBS = AAGGGAGGAG][linker = $(N)_{1-100}$][start codon and template-specific sequence = $ATG(N)_{0-100}$]3'.

Analogously, in a relevant, but by no means limiting example, an oligo can be used to provide a preferred 3' terminal region that is serviceable for topoisomerase I-based end-selection, which oligo is comprised of: a 1-10 base sequence that is convertible into a sticky end (preferably by a vaccinia topoisomerase I), and optional linker sequence followed by a template-specific sequence of 0-100 bases (to facilitate annealment to the template in the polymerase-based reaction). Thus, according to this example, a serviceable oligo (which may be termed a reverse primer) can have the sequence: 5'[terminal sequence = $(N)_{1-10}$][topoisomerase I site = AAGGG][linker = $(N)_{1-100}$][template-specific sequence = $(N)_{0-100}$]3'.

It is appreciated that, end-selection can be used to distinguish and separate parental template molecules (e.g. to be subjected to mutagenesis) from progeny molecules (e.g. generated by mutagenesis). For example, a first set of primers, lacking in a topoisomerase I recognition site, can be used to modify the terminal regions of the parental molecules (e.g. in polymerase-based amplification). A different second set of primers (e.g. having a topoisomerase I recognition site) can then be used to generate mutated progeny molecules (e.g. using any polynucleotide chimerization method, such as interrupted synthesis, template-switching polymerase-based amplification, or interrupted synthesis; or using saturation mutagenesis; or using any other method for introducing a topoisomerase I recognition site into a mutagenized progeny molecule as disclosed herein) from the amplified template molecules. The use of topoisomerase I-based end-selection can then facilitate, not only discernment, but selective topoisomerase I-based ligation of the desired progeny molecules.

Annealment of a second set of primers to thusly amplified parental molecules can be facilitated by including sequences in a first set of primers (i.e. primers used for amplifying a set parental molecules) that are similar to a toposiomerase I recognition site, yet different enough to prevent functional toposiomerase I enzyme recognition. For example, sequences that diverge from the AAGGG site by anywhere from 1 base to all 5 bases can be incorporated into a first set of primers (to be used for amplifying the parental templates prior to subjection to mutagenesis). In a specific, but non-limiting aspect, it is thus provided that a parental molecule can be amplified using the following exemplary – but by no means limiting – set of forward and reverse primers:

Forward Primer: 5' CTAGAAGAGAGGAGAAAACCATG$(N)_{10\text{-}100}$ 3', and
Reverse Primer: 5' GATCAAAGGCGCGCCTGCAGG$(N)_{10\text{-}100}$ 3'

According to this specific example of a first set of primers, $(N)_{10\text{-}100}$ represents preferably a 10 to 100 nuoleotide-long template-specific sequence, more preferably a 10 to 50 nucleotide-long template-specific sequence, more preferably still a 10 to 30 nucleotide-long template-specific sequence, and even more preferably still a 15 to 25 nucleotide-long template-specific sequence.

According to a specific, but non-limiting aspect, it is thus provided that, after this amplification (using a disclosed first set of primers lacking in a true topoisomerase I recognition site), amplified parental molecules can then be subjected to mutagenesis using one or more sets of forward and reverse primers that do have a true topoisomerase I recognition site. In a specific, but non-limiting aspect, it is thus provided that a parental molecule can be used as templates for the generation of a mutagenized progeny molecule using the following exemplary – but by no means limiting – second set of forward and reverse primers:

Forward Primer: 5' CTAGAAGGGAGGAGAAAACCATG 3'
Reverse Primer: 5' GATCAAAGGCGCGCCTGCAGG 3' (contains *Asc* I recognition sequence)

It is appreciated that any number of different primers sets not specifically mentioned can be used as first, second, or subsequent sets of primers for end-selection consistent with this invention. Notice that type II restriction enzyme sites can be incorporated (e.g. an *Asc* I site in the above example). It is provided that, in addition to the other sequences mentioned, the experimentalist can incorporate one or more N,N,G/T triplets into a serviceable primer in order to subject a working polynucleotide to saturation mutagenesis. Summarily, use of a second and/or subsequent set of primers can achieve dual goals of introducing a topoisomerase I site and of generating mutations in a progeny polynucleotide.

Thus, according to one use provided, a serviceable end-selection marker is an enzyme recognition site that allows an enzyme to cleave (including nick) a polynucleotide at a specified site, to produce a ligation-compatible end upon denaturation of a generated single stranded oligo. Ligation of the produced polynucleotide end can then be accomplished by the same enzyme (e.g. in the case of vaccinia virus topoisomerase I), or alternatively with the use of a different enzyme. According to one aspect of this invention, any serviceable end-selection markers, whether like (e.g. two vaccinia virus topoisomerase I recognition sites) or unlike (e.g. a class II restriction enzyme recognition site and a vaccinia virus topoisomerase I recognition site) can be used in combination to select a polynucleotide. Each selectable polynucleotide can thus have one or more end-selection markers, and they can be like or unlike end-selection markers. In a particular aspect, a plurality of end-selection markers can be located on one end of a polynucleotide and can have overlapping sequences with each other.

It is important to emphasize that any number of enzymes, whether currently in existence or to be developed, can be serviceable in end-selection according to this invention. For example, in a particular aspect of this invention, a nicking enzyme (e.g. N. *Bst*NB I, which cleaves only one strand at 5'...GAGTCNNNN/N...3') can be used in conjunction with a source of polynucleotide-ligating activity in order to achieve end-selection. According to this embodiment, a recognition site for N. *Bst*NB I – instead of a recognition site for topoisomerase I – should be incorporated into an end-selectable polynucleotide (whether end-selection is used for selection of a mutagenized progeny molecule or whether end-selection is used apart from any mutagenesis procedure).

It is appreciated that the instantly disclosed end-selection approach using topoisomerase-based nicking and ligation has several advantages over previously available selection methods. In sum, this approach allows one to achieve direction cloning (including expression cloning). Specifically, this approach can be used for the achievement of: direct ligation (i.e. without subjection to a classic restriction-purification-ligation reaction, that is susceptible to a multitude of potential problems from an initial restriction reaction to a ligation reaction dependent on the use of T4

DNA ligase); separation of progeny molecules from original template molecules (e.g. original template molecules lack topoisomerase I sites that not introduced until after mutagenesis), obviation of the need for size separation steps (e.g. by gel chromatography or by other electrophoretic means or by the use of size-exclusion membranes), preservation of internal sequences (even when topoisomerase I sites are present), obviation of concerns about unsuccessful ligation reactions (e.g. dependent on the use of T4 DNA ligase, particularly in the presence of unwanted residual restriction enzyme activity), and facilitated expression cloning (including obviation of frame shift concerns). Concerns about unwanted restriction enzyme-based cleavages – especially at internal restriction sites (or even at often unpredictable sites of unwanted star activity) in a working polynucleotide – that are potential sites of destruction of a working polynucleotide can also be obviated by the instantly disclosed end-selection approach using topoisomerase-based nicking and ligation.

3.3 Tunable

3.4 Transposons

3.4.1. GENERAL APPLICATIONS

In one aspect, the present invention relates generally to the field of transposable nucleic acid and for introducing genetic changes to nucleic acid. In one embodiment this invention relates to transposable elements isolated from maize and a process for using the same to identify and isolate genes and to insert desired gene sequences into plants in a heritable manner. In another embodiment, this invention provides for using transposons as a high molecular weight cloning system.

3.4.2. SPECIFIC METHODOLOGIES

3.4.2.1. Description Of Transposable Elements

Transposable genetic elements are DNA sequences, found in a wide variety of prokaryotic and eukaryotic organisms, that can move or transpose from one position to another position in a genome. In vivo, intra-chromosomal transpositions as well as transpositions between chromosomal and non-chromosomal genetic material are known. In several systems, transposition is known to be under the control of a transposase enzyme that is typically encoded by the transposable element. The genetic structures and transposition mechanisms of various transposable elements are summarized, for example, in "Transposable Genetic Elements" in "The Encyclopedia of Molecular Biology," Kendrew and Lawrence, Eds., Blackwell Science, Ltd., Oxford (1994), incorporated herein by reference.

Scientists have taken advantage of transposons to transport reporter genes for use in studying gene expression. These include transcriptional (Type I) fusions and translational (Type II) fusions. Transcriptional fusions, unlike translational fusions, place a reporter gene under the control of another promoter, but do not translationally fuse two protein domains. Translational fusions have generally been made to link a reporter gene carried inside the transposon to the translational frame of the target gene so that the reporter gene is expressed under direct control of the transcription and translation signals of the target gene of interest to study gene regulation. This requires that an open reading frame extend through the end of the transposable element to join an internal reporter protein to external translational sequences. This usually results in complete inactivation of the target gene.

3.4.2.2. In Vitro Transposition Systems

In vitro transposition systems that utilize the particular transposable elements of bacteriophage Mu and bacterial transposon Tn10 have been described, by the research groups of Kiyoshi Mizuuchi and Nancy Kleckner, respectively.

The bacteriophage Mu system was first described by Mizuuchi, K., "In Vitro Transposition of Bacteria Phage Mu: A Biochemical Approach to a Novel Replication Reaction," Cell:785-794 (1983) and Craigie, R. et al., "A Defined System for the DNA Strand-Transfer Reaction at the Initiation of Bacteriophage Mu Transposition: Protein and DNA Substrate Requirements," P.N.A.S. U.S.A. 82:7570-7574 (1985). The DNA donor substrate (mini-Mu) for Mu in vitro reaction normally requires six Mu transposase binding sites (three of about 30 bp at each end) and an enhancer sequence located about 1 kb from the left end. The donor plasmid must be supercoiled. Proteins required are Mu-encoded A and B proteins and host-encoded HU and IHF proteins. Lavoie, B. D, and G. Chaconas, "Transposition of phage Mu DNA," Curr. Topics Microbiol. Immunol. 204:83-99 (1995). The Mu-based system is disfavored for in vitro transposition system applications because the Mu termini are complex and sophisticated and because transposition requires additional proteins above and beyond the transposase.

The Tn10 system was described by Morisato, D. and N. Kleckner, "Tn10 Transposition and Circle Formation in vitro," Cell 51:101-111 (1987) and by Benjamin, H. W. and N. Kleckner, "Excision Of Tn10 from the Donor Site During Transposition Occurs By Flush Double-Strand Cleavages at the Transposon Termini," P.N.A.S. U.S.A. 89:4648-4652 (1992). The Tn10 system involves a supercoiled circular DNA molecule carrying the transposable element (or a linear DNA molecule plus *E. coli* IHF protein). The transposable element is defined by complex 42 bp terminal sequences with IHF binding site adjacent to the inverted repeat. In fact, even longer (81 bp) ends of Tn10 were used in reported experiments. Sakai, J. et al., "Identification and Characterization of Pre-Cleavage Synaptic Complex that is an Early Intermediate in Tn10 transposition," E.M.B.O. J. 14:4374-4383 (1995). In the Tn10 system, chemical treatment of the transposase protein is essential to support active transposition. In addition, the termini of the Tn10 element limit its utility in a generalized in vitro transposition system.

Both the Mu- and Tn10-based in vitro transposition systems are further limited in that they are active only on covalently closed circular, supercoiled DNA targets. What is desired is a more broadly applicable in vitro transposition system that utilizes shorter, more well defined termini and which is active on target DNA of any structure (linear, relaxed circular, and supercoiled circular DNA).

According to alternative embodiments of this invention, the steps of introducing a plurality of traits and/or generating a set of mutagenized organisms may include the step of cloning. In a preferred embodiment, this invention provides that the step of cloning may comprise using a Tn7 transposon-based system, such as but not limited to GPS-1. GPS-1 is an *in vitro* system (New England BioLabs Inc., Catalog #E7100S) that uses TnsABC* Transposase to insert a transposon (Transprimer™) randomly into the DNA target (See references Craig, N. L. (1996) Curr Top Microbiol Immunol 204, 27-48; Stellwagen, A. E. and Craig, N. L. (1997) Genetics 145, 573-85; Biery, M. C., Stewart, F. J., Stellwagen, A. E., Raleigh, E. A. and Craig, N. L., (2000) Nucleic Acids Res 28, 1067-1077). Such a system or modifications thereof that take advantage of transposon insertion sequences can be utilized to aid in the cloning of high molecular weight DNA. Such cloning approaches may also be provided for by this invention as a step in cell screening.

3.4.2.3. Importance Of Transposons In Agriculture

Currently, there is a great deal of interest in the development of gene transfer vectors for use with agriculturally important plants (See Outlook for Science and Technology, The Next Five Years, Vol. III (National Science Foundation (1982); and O.T.A. Report, Impact of Applied Genetics (1981)).

Although the United States presently has an excess productivity in the agricultural sector, this is recognized as a local and short term condition. Thus, agricultural research and planning must be based on long term considerations. The variety of problems surrounding increases in population, degradation of prime farm land and decreasing availability of good farm land necessitates the increased use of marginal land, as well as exogenous fertilizers and chemical pest control compositions.

Classical plant breeding programs have thus far been successful in increasing agricultural productivity. However, a substantial fraction of the increase in farm productivity experienced in the United States in the past 40 years is attributable to the use of fertilizers and modern energy intensive cultivation practices, both of which are increasingly costly. The ability of plant breeding alone to sustain productivity is a matter of some question. Plant breeders are divided in their views on whether genetic improvements will continue at the rate that has occurred over the past few decades or will begin to level out. Since such questions cannot be resolved a priori, it is prudent to explore a variety of additional means by which agronomically useful traits can be accumulated and improved in major crop plants. The unconventional areas that are presently receiving the most attention in the academic research establishment, as well as in both small and large firms with plant-oriented research programs are wide genetic crosses, tissue culture and the development of gene transfer systems that circumvent fertility barriers.

In the past, many attempts have been made to transform plant cells with DNA from a variety of sources. The first unequivocal demonstration that DNA transfer can and does occur in plants emerged from the work described above on *Agrobacterium tumefaciens* Ti plasmid. However, Ti-plasmid mediated gene transfer is presently accomplished only in dicotyledonous plants that interact with the plasmid's natural host bacterium. Since most major crop species are monocotyledonous, ti-plasmid mediated gene transfer has limited applications.

3.4.2.3.1. Use Of Transposons On The Ti Plasmid Of Agrobacterium

In higher organisms, transposons have been, or are being, used in several ways. For example, transposons are used as mutagens on the Ti plasmid of *Agrobacterium tumefaciens*. That is, a method for using bacterial transposons to cause insertion mutations in the Agrobacterium tumefaciens Ti plasmid, the causative agent of crown gall disease in dicotyledonous plants, has been developed. (See Zambriski, P., Goodman, H., Van Montagu, M. and Schell, J., Mobile Genetic Elements, J. Shapiro, Ed., (Academic Press) New York, pp. 506-535 (1983)). By this technique, it has become possible to identify the plasmid-borne genes that are responsible for virulence, as well as those that are responsible for the tumorous transformation of plant cells caused by the Ti plasmid. Further, it has become possible to show by using transposable elements, that a portion of the Ti plasmid can be integrated into plant genomes and can act as a vehicle for transferring genes from virtually any organism to any dicotyledonous plant that is susceptible to *Agrobacterium tumefaciens*.

3.4.2.3.2. Use Of Transposons In Maize

In maize, a monocotyledon, transposable elements were first genetically identified in the mid-1940s. These elements have been studied extensively and their genetic behavior has been extensively reviewed (See McClintock, B., Cold Spring Harbor Symp. Quant. Biol. 16:13-47 (1951); McClintock, B., Cold Spring Harbor Symp. Quant. Biol. 21:197-216 (1956); McClintock, B., Brookhaven Symp. Biol. 18:162-184 (1965); Fincham, J. R. S., and Sastry, G. R. K., Ann. Rev. Genet. 8:15-50 (1974); and Fedoroff, N., Mobile Genetic Elements, J. Shapiro, Ed., (Academic Press) New York, pp. 1-63 (1983)).

It has been demonstrated that transposons are normal, although cryptic, residents of the maize genome and that upon activation, they are responsible for various types of genetic rearrangements, including chromosome breakage, deletions, duplications, inversions and translocations. In addition, it has been shown that certain common types of unstable mutations, which have been studied for decades in both maize and in other organisms, are attributable to the insertion of transposons into genes or genetic loci.

3.4.2.4.1. Type I Fusions

Type I transcriptional fusions have been used to study gene expression and regulation by co-opting the native transcriptional signal to express the exogenous reporter gene. For example for gene expression in *E. coli,* yeast, and Drosophila development.

3.4.2.4.2. Type II Fusions

Type II fusions have also been used to study gene expression and regulation, but in this case not only co-opt the transcriptional signals, but any translational signals as well to express the reporter gene. In this type of system the protein product usually only expresses the activity of the reporter exogenous gene.

MudII elements are mini-Mu deletion elements which are type II Mu transposable elements. Examples of these include beta-galactosidase fusion elements, where a beta-galactosidase (lacZ) reporter gene is inserted via transposable elements to detect transcription and translation of regulated gene systems. This usually results in the inactivation of the targeted gene.

Two types of Mu protein fusions have been developed, lacZ fusion elements and nptI fusion elements (Symonds, Toussaint et al. (1987). Phage Mu) The lacZ elements have been used to study translation regulation, determination of the translation phase of target genes, infer the location of a protein fusion by hybrid protein size, determine amino terminal sequence, and raise antibodies to regions of the protein of interest. By far the major goal of these studies has been to determine mechanisms of gene expression in the studied organisms.

The nptI system was designed to perform transposon-tagging since nptI is known to function as an aminoglycoside resistance gene in a variety of organisms. Transposon tagging is a method of creating an mutant by inserting a transposon with a selectable marker into the gene of interest so that mutants which inactivate the gene can be identified and maintained. This element is useful since it allows the nptI to be directly linked to the transcription/translation system of the organism being studied. In these studies there has been no emphasis on creating novel proteins with new activities using these transposable elements. More importantly, these Mu elements are restricted to making amino-terminal fusions to the reporter protein. In these cases the inserted reporter gene is fused to the carboxy-end of the truncated targeted protein, terminating inside the Mu. If the transposable element were to insert before the amino terminal of a targeted gene, functional translation could only occur on the marker gene by itself, and no translation of the target gene would occur.

3.4.2.4.3. Problems with Mu

Unfortunately, available Mu elements had several problems. First, it has not been demonstrated that Mu elements can be readily used as a general method for the development of fusion proteins with two active domains. Second, the Mu elements used thus far for creation of protein fusions can not be used for construction of "carboxy-terminal" fusions since they did not have an open reading frame extending into the element. Third, the Mu elements previously used have long linker regions which incorporate a 40 amino acid linker between the fused domains. This could create protein folding problems or unwanted domain interactions. Fourth the currently existing Mu elements had only a single restriction site for the insertion of protein domains. Finally, although Mu elements which had deleted ends existed, it was not known whether they would transpose well with additional sequences added in such close proximity to the right end and whether the intervening linker region which would join the two protein domains would interfere with the construction of active chimeric proteins.

3.4.2.5. Other transposons

Other transposons have been used in a similar manner as Mu to create lac fusions to study gene expression. These include Tn10 and Tn917 (Berg and Howe. (1989). Mobile DNA).

The Tn5 element has also been used to construct phoA fusions in vivo. Fusions with alkaline phosphatase (phoA) have also been used to probe the structure of membrane bound proteins (Lloyd and Kadner. (1990). J Bacteriol. 172: 1688-93.).

In general, these transposons have been used to study the membrane topology structure of a particular gene and protein secretion. The resultant fusion proteins are also limited to amino-terminal fusion of the reporter PhoA reporter protein resulting in fusion at the carboxy end of the targeted gene.

In general, these types of fusions have been applied to the study of gene expression. These elements were constructed with truncated marker proteins that extend through the end of the transposon. Transposition of the element can create an in-frame fusion with a target gene, thereby activating expression. Mini-Mu elements are used because they transpose at high frequencies, insert randomly, and can be packaged along with a target plasmid and transduced to a new cell (Symonds, Toussaint et al. (1987). Phage Mu). Some of the more pertinent work that has been done in the area of transposable elements are detailed in the following.

Namgoong et al., (1994), teach that the Mu transposition reaction attachment sites attL and attR can promote the assembly of higher order complexes held together by non-covalent protein-DNA and protein-protein interactions. (Namgoong, Jayaram et al. (1994). J Mol Biol. 238: 514-527.)

Harel et al., (1990), teach that in Mu helper-mediated transposition packaging the left end contains an essential domain defined by nucleotides 1 to 54 of the left end (attL). At the right end (attR), they teach that the essential sequences for transposition require not more than the first 62 base pairs (bp), although the presence of sequences between 63 and 117 bp from the right end increase transposition frequency about 15-fold. (Harel, Dupliessis et al. (1990). Arch Microbiol. 154: 67-72.)

Groenen and van de Putte (1986), teach that the Mu A protein binds weakly to sequences between nucleotides 1 to 30 on the right end (R1) and between nucleotides 110 and 135 on the left end (L2). Mutations in these weak A binding sites have a greater effect on transposition than mutations of corresponding base pairs in the stronger A binding sites, located adjacent to these weak A binding sites. (Groenen and van de Putte. (1986). J Mol Biol. 189: 597-602.)

Groenen and et al. (1985) teach the DNA sequences at the end of the genome of bacteriophage Mu that are essential for transposition. (Groenen, Timmers et al. (1985). Proc Natl Acad Sci, USA. 82: 2087-2091.)

Lloyd and Kadner teach the how to probe the topology of the uhpT sugar phosphate transporter using a Tn5phoA element. (Lloyd and Kadner. (1990). J Bacteriol. 172: 1688-93.)

Phage Mu (1987), Cold Spring Harbor Laboratory Press (Symonds, et al eds.) teaches general methods for handling and working with bacteriophage Mu as a transposon, and describes the various uses of mini- Mu elements including the construction of Mu transcriptional and translational fusions.

Silhavy and Beckwith (1985) teaches the various uses of lac fusions for the study of biological problems. (Silhavy and Beckwith. (1985). Microbiol Rev. 49: 398-418.)

Mobile DNA, (1989), American Society for Microbiology, Publishers. (Berg, Howe, eds) describes transposons.

Casadaban, et al. (1983) Methods in Enzymol, provides a good general review of beta-galactosidase gene fusions for the study of gene expression. (Casadaban, Martinez-Arias et al. (1983). Recombinant DNA. Methods in Enzymology. 100: 293-308.)

3.4.3. In Vitro Transposition System

The present invention is summarized in that an in vitro transposition system comprises a preparation of a suitably modified transposase of bacterial transposon Tn5, a donor DNA molecule that includes a transposable element, a target DNA molecule into which the transposable element can transpose, all provided in a suitable reaction buffer.

3.4.3.1. Donor DNA Molecule: Transposable DNA Sequence Of Interest

The transposable element of the donor DNA molecule is characterized as a transposable DNA sequence of interest, the DNA sequence of interest being flanked at its 5'-and 3'-ends by short repeat sequences that are acted upon in trans by Tn5 transposase.

3.4.3.1.1. Modified Transposase Enzyme Comprises Two Classes Of Differences From Wild Type Tn5 Transposase

The invention is further summarized in that the suitably modified transposase enzyme comprises two classes of differences from wild type Tn5 transposase, where each class has a separate measurable effect upon the overall transposition activity of the enzyme and where a greater effect is observed when both modifications are present. The suitably modified enzyme both (1) binds to the repeat sequences of the donor DNA with greater avidity than wild type Tn5 transposase ("class (1) mutation") and (2) is less likely than the wild type protein to assume an inactive multimeric form ("class (2) mutation"). A suitably modified Tn5 transposase of the present invention that contains both class (1) and class (2) modifications induces at least about 100-fold (.+/-.10%) more transposition than the wild type enzyme, when tested in combination in an in vivo conjugation assay as described by Weinreich, M. D., "Evidence that the cis Preference of the Tn5 Transposase is Caused by Nonproductive Multimerization," Genes and Development 8:2363-2374 (1994), incorporated herein by reference. Under optimal conditions, transposition using the modified transposase may be higher. A modified transposase containing only a class (1) mutation binds to the repeat sequences with sufficiently greater avidity than the wild type Tn5 transposase that such a Tn5 transposase induces about 5- to 50-fold more transposition than the wild type enzyme, when measured in vivo. A modified transposase containing only a class (2) mutation is sufficiently less likely than the wild type Tn5 transposase to assume the multimeric form that such a Tn5 transposase also induces about 5- to 50-fold more transposition than the wild type enzyme, when measured in vivo.

3.4.4. Transposons - Specialized Applications

3.4.4.1. In Vitro System For Introducing Any Transposable Element From A Donor DNA Into A Target DNA

It will be appreciated that this technique provides a simple, in vitro system for introducing any transposable element from a donor DNA into a target DNA. It is generally accepted and understood that Tn5 transposition requires only a pair of OE termini, located to either side of the transposable element. These OE termini are generally thought to be 18 or 19 bases in length and are inverted repeats relative to one another. Johnson, R. C., and W. S. Reznikoff, Nature 304:280 (1983), incorporated herein by reference. The Tn5 inverted repeat sequences, which are referred to as "termini" even though they need not be at the termini of the donor DNA molecule, are well known and understood.

Apart from the need to flank the desired transposable element with standard Tn5 outside end ("OE") termini, few other requirements on either the donor DNA or the target DNA are envisioned. It is thought that Tn5 has few, if any, preferences for insertion sites, so it is possible to use the system to introduce desired sequences at random into target DNA. Therefore, it is believed that this method, employing the modified transposase described herein and a simple donor DNA, is broadly applicable to introduce changes into any target DNA, without regard to its nucleotide sequence.

3.4.4.2. Generation of functional fusion protein products

The instant invention provides constructs and methods for the rapid and efficient generation of functional fusion protein products with either carboxy-terminal or amino-terminal fusions. Functional fusion proteins are those which retain some of the activity of the original domains, and/or those which have a newly created activity. Throughout this specification, reference is made to two types of fusions: carboxy terminal fusions and amino terminal fusions. In this text we use amino and carboxy terminal fusions to refer to the end of the domain inside of the Mu elements which is fused to the target molecule. Thus, carboxy terminal fusion elements are those with a protein domain inside of the Mu which extends out of the Mu element such that the exogenous protein is fused to the amino end of the endogenous protein. The amino terminal fusion elements are those that create fusions with a target gene extending into the element such that the exogenous protein is fused to the carboxy terminal of the endogenous protein (as referenced within U.S. Patent Numbers 5,965,443 and 4,732,856).

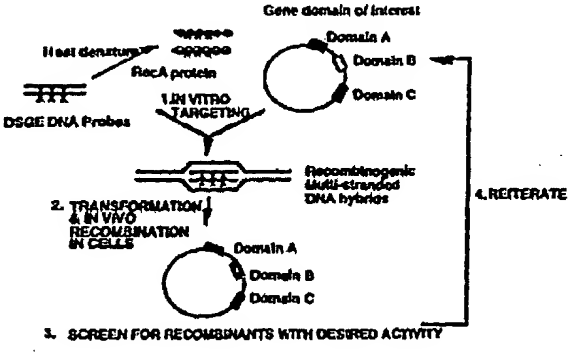
3.4.5. Additional Applications

It is envisioned that in addition to the uses specifically noted herein, other applications will be apparent to the skilled molecular biologist. In particular, methods for introducing desired mutations into prokaryotic or eukaryotic DNA are very desirable. For example, at present it is difficult to knock out a functional eukaryotic gene by homologous recombination with an inactive version of the gene that resides on a plasmid. The difficulty arises from the need to flank the gene on the plasmid with extensive upstream and downstream sequences. Using this system, however, an inactivating transposable element containing a selectable marker gene (e.g., neo) can be introduced in vitro into a plasmid that contains the gene that one desires to inactivate. After transposition, the products can be introduced into suitable host cells. Using standard selection means, one can recover only cell colonies that contain a plasmid having the transposable element. Such plasmids can be screened, for example by restriction analysis, to recover those that contain a disrupted gene. Such clones can then be introduced directly into eukaryotic cells for homologous recombination and selection using the same marker gene.

Also, one can use the system to readily insert a PCR-amplified DNA fragment into a vector, thus avoiding traditional cloning steps entirely. This can be accomplished by (1) providing suitable a pair of PCR primers containing OE termini adjacent to the sequence-specific parts of the primers, (2) performing standard PCR amplification of a desired nucleic acid fragment, (3) performing the in vitro transposition reaction of the present invention using the double-stranded products of PCR amplification as the donor DNA.

The invention is not intended to be limited to the foregoing examples, but to encompass all such modifications and variations as come within the scope of the appended claims.

3.5. Homologous Recombination



3.5.1. Homologous Recombination For The Generation Of Deletions And Insertions

The invention relates to compositions and methods of rapidly evolving specific protein domains using a library of nucleic acid filaments and a recombinase polypeptide or peptide. The invention relates to compositions and methods for targeting sequence modifications in one or more genes of a related family of genes using enhanced homologous recombination techniques. The invention also relates to compositions and methods for isolating and identifying novel members of homologous sequences families. These techniques may be used to create animal or plant models of disease as well as to identify new targets for drug or pathogen screening.

3.5.1.1. Evolution Of Genes

In nature, the evolution of genes and their encoded proteins occurs through an equilibrium between recombination or mutation and selection. While evolution in nature takes millions of years, in vitro methods and compositions have been developed to evolve proteins, with improved and novel functions, in a matter of hours to days.

3.5.1.1.1. Through Mutagenesis

Current in vitro gene evolution methods utilize repeated cycles of random mutagenesis or random nicking and mixing of related genes containing mutations in PCR-based random recombination. These methods couple multiple rounds of in vitro mutagenesis with screening systems to produce and identify the desired mutants or recombinants (Stemmer 1994. Nature 370:389-391; Arnold 1996. Chemical Engineering Science 51:5091-5102). Research has shown, however, that the mutations of interest tend to occur in those regions or domains that are directly related to function (Chen and Arnold. 1993. PNAS USA 90:5618-5622).

However, these mutagenesis methods produce random mutations throughout the gene of interest which requires the need to screen large numbers of uninteresting or deleterious mutants. The labor-intensive and time consuming aspects of these methods are further complicated by the necessity of multiple rounds of subcloning and can be extremely challenging if the screening system is complex and does not utilize a selection system.

3.5.1.1.2. Homologous Recombination (HR)

Homologous recombination (HR) is defined as the exchange of homologous or similar DNA sequences between two DNA molecules. As essential feature of HR is that the enzymes responsible for the recombination event can pair any homologous sequences as substrates. The ability of HR to transfer genetic information between DNA molecules makes targeted homologous recombination a very powerful method in genetic engineering and gene manipulation. Both genetic and cytological studies have indicated that such a crossing-over process occurs between pairs of homologous chromosomes during meiosis in higher organisms.

3.5.1.1.3. Site-Specific Recombination

Alternatively, in site-specific recombination, exchange occurs at a specific site, as in the integration of phage A into the E coli chromosome and the excision of lambda DNA from it. Site-specific recombination involves specific inverted repeat sequences; e.g. the Cre-loxP and FLP- FRT systems. Within these sequences there is only a short stretch of homology necessary for the recombination event, but not sufficient for it. The enzymes involved in this event generally cannot recombine other pairs of homologous (or nonhomologous) sequences, but act specifically.

3.5.1.1.4. Advantage Of Homologous Recombination Over Site-Specific Recombination

Although both site-specific recombination and homologous recombination are useful mechanisms for genetic engineering of DNA sequences, targeted homologous recombination provides a basis for targeting and altering essentially any desired sequence in a duplex DNA molecule, such as targeting a DNA sequence in a chromosome for replacement by another sequence. Site- specific recombination has been proposed as one method to integrate transfected DNA at chromosomal locations having specific recognition sites (O'Gorman et al. (1991) Science 251: 1351; Onouchi et al. (1991) Nucleic Acids Res. 19: 6373). Unfortunately, since this approach requires the presence of specific target sequences and recombinases, its utility for targeting recombination events at any particular chromosomal location is severely limited in comparison to targeted general recombination.

3.5.1.1.5. HR To Create Transgenic Plants, Animals, And Organisms

Homologous recombination has also been used to create transgenic plants and animals. Transgenic organisms contain stably integrated copies of genes or gene constructs derived from another species in the chromosome of the transgenic organism. In addition, gene targeted animals can be generated by introducing cloned DNA constructs of the foreign genes into totipotent cells by a variety of methods, including homologous recombination. For example, animals that develop from genetically altered totipotent cells can contain the foreign gene in all somatic cells and also in germ-line cells.

3.5.1.1.5.1. Current Methods Using Embryonic Stem Cells

Currently methods for producing transgenic and targeted animals have been performed on totipotent embryonic stem cells (ES) and with fertilized zygotes. ES cells have an advantage in that large numbers of cells can be manipulated easily by homologous recombination in vitro before they are used to generate targeted animals. Currently, however, only embryonic stem cells from mice have been shown to contribute to the germ line. Alternatively, DNA can also be introduced into fertilized oocytes by micro-injection into pronuclei which are then transferred into the uterus of a pseudo-pregnant recipient animal to develop to term. The ability of mammalian and human cells to incorporate exogenous genetic material into genes residing on chromosomes has demonstrated that these cells have the general enzymatic machinery for carrying out homologous recombination required between resident and introduced sequences. These targeted recombination events can be used to correct mutations at known sites, replace genes or gene segments with defective ones, or introduce foreign genes into cells.

3.5.1.1.5.2. Frequency And Efficiency Of HR

HR can be used to add subtle mutations at known sites, replace wild type genes or gene segments or introduce completely foreign genes into cells. However, HR efficiency is very low in living cells and is dependent on several parameters, including the method of DNA delivery, how it is packaged, its size and conformation, DNA length and position of sequences homologous to the target, and the efficiency of hybridization and recombination at chromosomal sites. These variables severely limit the use of conventional HR approaches for gene evolution in cell based systems.

(Kucherlapati et al., 1984. PNAS; USA 81:3153-- 3157; Smithies et al. 1985. Nature 317:230-234; Song et al. 1987. PNAS USA 84:6820-6824; Doetschman et al. 1987. Nature 330:576-578; Kim and Smithies. 1988. Nuc. Acids. Res. 16:8887- 8903; Koller and Smithies. 1989. PNAS USA 86:8932-8935; Shesely et al. 1991. PNAS USA 88:4294- 4298; Kim et al. 1991. Gene 103:227-233).

3.5.1.1.5.2.1. Enhancement By The Presence Of Recombinase Activities

The frequency of HR is significantly enhanced by the presence of recombinase activities in cellular and cell free systems. Several proteins or purified extracts that promote HR (i.e., recombinase activity) have been identified in prokaryotes and eukaryotes (Cox and Lehman., 1987. Annu. Rev. Biochem. 56:229-262; Radding. 1982. Annual Review of Genetics 16:405-547; McCarthy et al. 1988. PNAS; USA 85:5854-5858). These recombinases promote one or more steps in the formation of homologously-paired intermediates, strand-exchange, and/or other steps.
Recent advances have resulted in techniques allowing enhanced homologous recombination (EHR) using recombinases such as recA and Rad5l and single-stranded nucleic acids that have sequence heterologies. This allows sequence modifications to be specifically targeted to virtually any genomic position. See for example, PCT US93/03868 and PCT US98/05223, both of which are expressly incorporated herein by reference.

3.5.1.1.5.2.1.1. Recombinase Rec A: A Bacterial Protein That Catalyses Homologous Pairing And Strand Exchange Between Two Homologous DNA Molecules

The most studied recombinase to date is the RecA recombinase of E coli, which is involved in homology search and strand exchange reactions (Cox and Lehman, 1987, supra). The bacterial RecA protein (Mr 37,842) catalyses homologous pairing and strand exchange between two homologous DNA molecules (Kowalczykowski et al. 1994. Microbiol. Rev. 58:401-465; West. 1992. Annu. Rev. Biochem. 61:603-640); Roca and Cox. 1990. CRC Cit. Rev. Biochem. Mol. Biol. :415-455; Radding. 1989. Biochim. Biophys. Acta. 1008:131-145; Smith. 1989. Cell 58:807-809).

RecA protein binds cooperatively to any given sequence of single- stranded DNA with a stochiometry of one RecA protein monomer for every three to four nucleotides in DNA (Cox and Lehman, 1987, supra). This forms unique right handed helical nucleoprotein filaments in which the DNA is extended by 1.5 times its usual length (Yu and Egelman 1992. J. Mol. Biol. 227:334-346). These nucleoprotein filaments, which are referred to as DNA probes, are crucial "homology search engines" which catalyze DNA pairing. Once the filament finds its homologous target gene sequence, the DNA probe strand invades the target and forms a hybrid DNA structure, referred to as a joint molecule or D-loop (DNA displacement loop) (McEntee et al. 1979. PNAS USA 76:2615-2619; Shibata et al. 1979. PNAS USA 76:1638-1642). The phosphate backbone of DNA inside the RecA nucleoprotein filaments is protected against digestion by phosphodiesterases and nucleases.

RecA protein is the prototype of a universal class of recombinase enzymes which promote probe-target pairing reactions. Recently, genes homologous to E.coli RecA (the Rad5l family of proteins) were isolated from all groups of eukaryotes, including yeast and humans. Rad5l protein promotes homologous pairing and strand invasion and exchange between homologous DNA molecules in a similar manner to RecA protein (Sung. 1994. Science 265:1241-1243; Sung and Robberson. 1995. Cell 82:453-461; Gupta et al. 1997. PNAS USA 94:463-468; Baumann et al. 1996. Cell 87:757-766).

3.5.1.1.5.3. Functional Genomics: The Correlation Of Genotype And Phenotype

One area of pressing interest in biology is within the area of "functional genomics", i.e. the correlation of genotype and phenotype. This requires animal systems, since phenotypic changes must be evaluated in vivo. Similarly, and related to this idea, is the elucidation and characterization of gene families, i.e. genes or proteins that are structurally related, i.e. they have sequence homologies between the members of the family. Since presumably many, if not most, disease states are caused by multiple gene interactions, the ability to evaluate interactions among genes, and particularly within or between gene families, at the phenotype level, would be extremely valuable.

The functional genomics tools that allow facile identification and engineering of gene family members in animals and cells, however, are not yet available. While the amino acid sequence motifs shared between gene family members may be identical, due to degeneracy in the DNA code, the DNA sequence identity may be significantly less. Hence, one criterion necessary for genetic modifications of gene family members is development of homologous recombination technologies that can be used to clone and modify similar DNA sequences that share little sequence identity. This is particularly important since homologous recombination in cells normally requires significant sequence identity to work efficiently. Relaxing the amount of sequence identity needed for homologous recombination allows greater flexibility to target related genes for creating transgenic animals and cells containing modifications in gene family consensus sequences, and also will allow the rapid cloning, generation of gene family specific libraries, and evolution of gene family members. Accordingly, it is an object of the invention to provide an efficient method of domain specific gene evolution that generates maximal diversity but increases the probability of identifying a gene of interest.

3.5.2. Domain Specific Gene Evolution

3.5.2.1. Domain Specific Gene Evolution - Comprising Forming A Plurality Of Recombination Intermediates Comprising A Target Nucleic Acid Encoding An Amino Acid Sequence Of Interest, A Recombinase And A Plurality Of Targeting Polynucleotides

The present invention provides methods of domain specific gene evolution comprising forming a plurality of recombination intermediates comprising a target nucleic acid encoding an amino acid sequence of interest, a recombinase and a plurality of targeting polynucleotides. The targeting polynucleotides are substantially complementary to each other and each comprises a homology clamp that substantially correspond to or is substantially complementary to a predetermined sequence of the target nucleic acid and comprise random or degenerate sequences. The predetermined sequence encodes a domain of the amino acid sequence. The method further comprises contacting the intermediate with a recombination proficient cell, whereby a library of altered target nucleic acids are produced. The altered target nucleic acids are expressed in the cell to generate a pool of variant amino acid sequences. The method further comprises selecting and isolating a cell comprising an altered target nucleic acid that expressed a variant amino acid having a desired activity.

3.5.2.2. Comprising Forming A Recombination Intermediate Comprising A Target Nucleic Acid Encoding An Amino Acid Sequence Of Interest, A Recombinase And A Pair Of Targeting Poly Nucleotides

In another aspect of the invention, a method of domain specific gene evolution comprises forming a recombination intermediate comprising a target nucleic acid encoding an amino acid sequence of interest, a recombinase and a pair of targeting polynucleotides. The targeting polynucleotides are substantially complementary to each other and each comprises a homology clamp that substantially corresponds to or is substantially complementary to a predetermined sequence of the target nucleic acid. The predetermined sequence encodes a domain of the amino acid sequence. The method further comprises contacting the intermediate with a single-strand specific nuclease or junction-specific nuclease to form a nicked or open-ended target nucleic acid. The regions adjacent to the hybridized region or junctions are susceptible to nucleases. The target nucleic acid is reassembled and recombined to produce a library of altered target nucleic acids. The target nucleic acids are expressed to generate a pool of variant amino acid sequences. The variant amino acid sequences are selected and characterized to identify an altered target nucleic acid encoding a variant amino acid sequence of interest.

In a further aspect, each method is repeated one or more times to further evolve a variant amino acid sequence having a desired activity. In yet another aspect, more than one domain or a protein is evolved simultaneously.

3.5.2.3. Compositions

It is an object of the present invention to provide compositions comprising at least one recombinase and at least two single-stranded targeting polynucleotides which are substantially complementary to each other and each having a consensus homology clamp for a gene family.

In an additional aspect, the invention provides compositions comprising at least one recombinase and a plurality of pairs of single stranded targeting polynucleotides, where the plurality of pairs comprises a set of degenerate probes encoding the consensus sequence.

In a further aspect, the invention provides kits comprising the compositions of the invention and at least one reagent.

3.5.2.4. Methods For Targeting A Sequence Modification In At Least One Member Of A Consensus Family Of Genes In A Cell By Homologous Recombination.

In an additional aspect, the invention provides methods for targeting a sequence modification in at least one member of a consensus family of genes in a cell by homologous recombination. The method comprises introducing into at least one cell at least one recombinase and at least two single-stranded targeting polynucleotides which are substantially complementary to each other and each having a consensus homology clamp for the family. The method can additionally comprise identifying a target cell having a targeted sequence modification.

3.5.2.4.1. Methods Of Making A Non- Human Organism With A Targeted Sequence Modification In At Least One Member Of A Gene Family

In a further aspect, the invention provides methods of making a non- human organism with a targeted sequence modification in at least one member of a gene family. The method comprises introducing into a cell at least one recombinase and at least two single-stranded targeting polynucleotides which are substantially complementary to each other and each having a consensus homology clamp for said family. The cell is then subjected to conditions that result in the formation of an animal, and the animal has at least one modification in at least one member of a consensus family of genes.

In a further aspect, the invention provides non-human organisms containing a sequence modification in an endogenous consensus functional domain of a gene member of a gene family.

3.5.2.5. Methods Of Isolating A Member Of A Gene Family Comprising A Protein Consensus Sequence

In an additional aspect, the invention provides methods of isolating a member of a gene family comprising a protein consensus sequence. The method comprises adding to a complex mixture of nucleic acids at least one recombinase and at least two single-stranded targeting polynucleotides which are substantially complementary to each other and each having a consensus homology clamp for said family. At least one of the targeting polynucleotides comprises a purification tag. The method is done under conditions whereby the targeting polynucleotides form a complex with the member, and the family member is isolated using said purification tag. The complex nucleic acid mixture may be a cDNA library, a cell, RNA or a restriction endonucleases genomic digest.

3.5.3. Targeting A Predetermined Nucleic Acid Sequence That Encodes A Specific Protein Domain, To Make A Plurality Of Targeted Sequence Modifications

The present invention provides methods and compositions for domain specific gene evolution. In one aspect of the invention, the method comprises targeting a predetermined nucleic acid sequence that encodes a specific protein domain, to make a plurality of targeted sequence modifications. That is, by targeting the recombinogenic probes of the invention to particular protein domains, gene evolution and selection are targeted to specific domains known or believed to harbor specific activities or functions. These methods create maximal diversity in specific domains of interest, thereby, decreasing the size of the library of mutations that are to be screened and increasing the probability of finding a gene with improved or desired attributes. Therefore, the libraries of the present invention are enriched for advantageous or interesting mutations or recombinant sequence(s).

3.5.3.1. Combining A Plurality Of Pairs Of Single-Stranded Targeting Poly Nucleotides, A Predetermined Target Nucleic Acid, And A Recombinase To Form A Polynucleotide:Target Nucleic Acid Complex

Accordingly, the methods comprise combining a plurality of pairs of single-stranded targeting poly nucleotides, a predetermined target nucleic acid, and a recombinase to form a polynucleotide:target nucleic acid complex. The targeting polynucleotides comprise at least one homology clamp for targeting a predetermined domain of a target nucleic acid and randomized or degenerate sequences. The complex is optionally introduced into a plurality of recombination proficient cells which catalyze strand exchange and homologous recombination intracellularly to produce a library of modified nucleic acids. Cells are selected and isolated that comprise a modified nucleic acid that encodes a polypeptide having a desired property. The process is preferably repeated iteratively to further evolve the target domain of interest.

3.5.3.2. Domain Specific DNA Nicking

In another aspect of the invention, methods of domain specific DNA nicking are provided for domain specific gene evolution. This method comprises combining a pair of single-stranded targeting polynucleotides, a predetermined target nucleic acid, and a recombinase to form a polynucleotide:target nucleic acid complex. The targeting polynucleotides are substantially complementary and comprise at least one homology clamp for targeting a predetermined domain of a target nucleic acid. The polynucleotide:target nucleic acid complex is treated with a single-strand specific nuclease, which preferentially nicks the regions flanking the polynucleotide:target nucleic acid complex region (Ferrin and Camerini-Otero. 1991. Science. 254 1494-1497). That is, the domain is protected from recombination by the initial presence of the recombinase in the complex. The nuclease is inactivated and the complex dissociated. The nicked target nucleic acid is reassembled and recombined by PCR to produce a library of nucleic acids with preferential modifications in the nicked regions. The library of modified nucleic acids can be introduced into a host cell and expressed. Cells are selected and isolated that comprise a modified nucleic acid that encodes a polypeptide having a desired property. This process is repeated iteratively to further evolve the predetermined targeted domain of interest.

In each of the methods described above, single domains and optionally multiple domains are targeted. The methods and compositions described above are optionally used in combination for domain specific gene evolution. For example, individual or multiple rounds of domain specific DNA nicking are followed or interspersed with one or more rounds of domain specific evolution employing a plurality of targeting polynucleotides described above.

The methods of the present invention also avoid multiple subcloning steps. This is particularly relevant when large complex vectors such as lambda, BACS, PACS, YACS, MACS and other genomic DNAs are used and where multiple subcloning steps make mutagenesis and shuffling of unique sites in large vectors particularly tedious and time consuming.

3.5.3.3. Generating Homologous Recombination Intermediates In Vitro, Panels Or Libraries Of Mutagenized And Shuffled Genes To Generate In Vitro Evolution

Accordingly, the present invention provides methods to introduce recombinogenic probe or hybrid complexes into recombination proficient cells to link in vitro and in vivo recombination and evolution processes. By generating homologous recombination intermediates in vitro, panels or libraries of mutagenized and shuffled genes are generated for in vitro evolution. The link to in vivo systems allows in vivo selection of evolved genes encoding proteins of a desired characteristic. The present invention can thus be used in a variety of important ways.

3.5.3.3.1. Methods Can Be Used In The Creation Of Transgenic Organisms, Animal, And Plant Models Of Disease

First, these methods can be used in the creation of transgenic organisms, animal, and plant models of disease. Thus, for example, domain-specific targeting polynucleotides used in homologous recombination methods can generate animals that have a wide variety of mutations in a wide variety of functionally related genes, potentially resulting in a wide variety of phenotypes, including phenotypes related to disease states. This may also be done on a cellular level, to identify genes involved in cellular phenotypes, i.e. target identification.

3.5.3.3.2. Identify "Reversion" Genes, Genes That Can Modulate Disease States

Secondly, domain targeting can be used in cells or animals that are diseased or altered; in essence, domain targeting can be done to identify "reversion" genes, genes that can modulate disease states caused by different genes, either genes within the same gene family or a completely different gene family. Thus, for example the loss of one type of enzymatic activity, resulting in a disease phenotype, may be compensated by alterations in a different but homologous enzymatic activity.

3.5.3.3.3. Creation Of Libraries Of Altered Nucleic Acids

In addition, the methods may be used in the creation of libraries of altered nucleic acids, including extrachromosomal sequences, and can be expressed in cells to produce libraries of altered proteins, which then can be screened for any number of useful or interesting properties, including, but not limited to, increased or altered stability (thermal, pH, oxidants, to proteases, etc.); altered specificity (for example, in the case of enzymes); altered binding; modified activity and other desirable properties, such as, altered immunogenicity.

3.5.3.4. Use Of Homology Motif Tags (HMTs) In Targeted Homologous Recombination To Elucidate Disease Mechanisms And To Identify Disease Targets Contained Within Gene Families

The present invention is directed to the use of homology motif tags (HMTs) in targeted homologous recombination to elucidate disease mechanisms and to identify disease targets contained within gene families related by the presence of one or more common domains. That is, there are a large number of gene families that contain genes related by the presence of similar functional domains, i.e. binding domains for substrates or other proteins, enzymatic domains such as kinase or protease domains, signaling and regulator domains, receptor binding domains, ATP binding domains, leucine zipper domains, zinc finger domains, etc. These functional domains frequently result in primary sequence homology; that is, related functional domains have related sequences. Many of these functional domains have been studied and so-called "consensus sequences" identified; that is, an average sequence derived from a number of related sequences. Each residue (or set of residues) of the consensus sequence is the most frequent at that position in the set under consideration. Consensus sequences can be either amino acid or nucleic acid consensus sequences, with amino acid sequences being used to generate nucleic acid consensus sequences.

Interestingly, while a wide variety of gene families are known, the majority of drug targets come from only four of these gene families. These are the G-protein coupled or seven-transmembrane domain receptors, nuclear (hormone) receptors, ion channels, esterases. Other important gene families are enzymes, including recombinases. Of the top 100 pharmaceutical drugs, 18 bind to seven- transmembrane receptors, 10 to nuclear receptors and 16 to ion channels.

By using HMTs directed to the consensus sequences of gene families for homologous recombination and particularly enhanced homologous recombination methods, sequence modifications may be made to any number of targeted genes in a related family.

3.5.3.4.1. Methods And Compositions Utilizing Homology Motif Tags (HMTs) Or Consensus Sequences

Accordingly, the present invention provides methods and compositions utilizing homology motif tags (HMTs) or consensus sequences. By "homology motif tag" or "protein consensus sequence" herein is meant an amino acid consensus sequence of a gene family. By "consensus nucleic acid sequence" herein is meant a nucleic acid that encodes a consensus protein sequence of a functional domain of a gene family. In addition, "consensus nucleic acid sequence" can also refer to cis sequences that are non-coding but can serve a regulatory or other role. As outlined below, generally a library of consensus nucleic acid sequences are used, that comprises a set of degenerate nucleic acids encoding the protein consensus sequence. A wide variety of protein consensus sequences for a number of gene families are known. A "gene family" therefore is a set of genes that encode proteins that contain a functional is domain for which a consensus sequence can be identified. However, in some instances, a gene family includes non-coding sequences; for example, consensus regulatory regions can be identified. For example, gene family/consensus sequences pairs are known for the G- protein coupled receptor family, the AAA-protein family, the bZIP transcription factor family, the mutS family, the recA family, the Rad51 family, the dmel family, the recF family, the SH2 domain family, the Bcl- 2 family, the single-stranded binding protein family, the TFIID transcription family, the TGF-beta family, the TNF family, the XPA family, the XPG family, actin binding proteins, bromodomain GDP exchange factors, MCM family, ser/thr phosphatase family, etc.

As will be appreciated by those in the art, the proteins of the gene families generally do not contain the exact consensus sequences; generally consensus sequences are artificial sequences that represent the best comparison of a variety of sequences. The actual sequence that corresponds to the functional sequence within a particular protein is termed a "consensus functional domain" herein; that is, a consensus functional domain is the actual sequence within a protein that corresponds to the consensus sequence. A consensus functional domain may also be a "predetermined endogenous DNA sequence" (also referred to herein as a "predetermined target sequence") that is a polynucleotide sequence contained in a target cell. Such sequences can include, for example, chromosomal sequences (e.g., structural genes, regulatory sequences including promoters and enhancers, recombinatorial hotspots, repeat sequences, integrated proviral sequences, hairpins, palindromes), episomal or extrachromosomal sequences (e.g., replicable plasmids or viral replication intermediates) including chloroplast and mitochondrial DNA sequences. By "predetermined" or "pre-selected" it is meant that the consensus functional domain target sequence may be selected at the discretion of the practitioner on the basis of known or predicted sequence information, and is not constrained to specific sites recognized by certain site-specific recombinases (e.g., FLP recombinase or CRE recombinase). In some embodiments, the predetermined endogenous DNA target sequence will be other than a naturally occurring germline DNA sequence (e.g., a transgene, parasitic, mycoplasmal or viral sequence).

3.5.3.4.1.1. Gene Family Is The G-Protein Coupled Receptor Family

3.5.3.4.1.1.1. Subfamily 1 Also Called R7G Proteins

In a preferred embodiment, the gene family is the G-protein coupled receptor family, which has over 900 identified members, including several subfamilies. In a preferred embodiment, the G-protein coupled receptors are from subfamily 1 and are also called R7G proteins. They are an extensive group of receptors which recognize hormones, neurotransmitters, odorants and light and transduce extracellular signals by interaction with guanine (G) nucleotide-binding proteins. The structure of all these receptors is thought to be virtually identical, and they contain seven hydrophobic regions, each of which putatively spans the membrane. The N-terminus is extracellular and is frequently glycosylated, and the C-terminus is cytoplasmic and generally phosphorylated. Three extracellular loops alternate with three cytoplasmic loops to link the seven transmembrane regions. G- protein coupled receptors include, but are not limited to: the class A rhodopsin first subfamily, including amine (acetylcholine (muscarinic), adrenoceptors, domamine, histamine, serotonin, octopamine), peptides (angiotensin, bombesin, bradykinin, C5a anaphylatoxin, Fmet-leu-phe, interleukin-8, chemokine, CCK, endothelin, mealnocortin, neuropeptide Y, neurotensin, opioid, somatostatin, tachykinin, thrombin, vasopressin-like, galanin, proteinase activated), hormone proteins (follicle stimulating hormone, lutropin-choriogonadotropic hormone, thyrotropin), rhodopsin (vertebrate), olfactory (olfactory type 1 -11, gustatory), prostanoid (prostaglandin, prostacyclin, thromboxane), nucleotide (adenosine, purinoceptors), cannabis, platelet activating factor, gonadotropin-releasing hormone (gonadotropin releasing hormone, thyrotropin-releasing hormone, growth hormone secretagogue), melatonin, viral proteins, MHC receptor, Mas proto-oncogene, EBV-induced and glucocorticoid induced; the class B secretin second subfamily, including calcitonin, corticotropin releasing factor, gastric inhibitory peptide, glucagon, growth hormone releasing hormone, parathyroid hormone, secretin, vasoactive intestinal polypeptide, and diuretic hormone; the class C metabotropic glutamate third subfamily, including metabrotropic glutamate and extracellular calcium-sensing agents; and the class D pheromone fourth subfamily. Because of the large number of family members, these large classes of GPCRs can be further subdivided into subfamilies. Examples of these subfamilies are calcitonin, glucagon, vasoactive and parathyroid are from class B; and acetylcholine, histamine angiotensin, alpha2- and beta-adrenergic are from class A. From each subfamily small protein consensus sequences can be derived from sequence alignments. For example, there are 6 motifs for the metabotripic glutamate like GPRCs derived from the indicated number of family members. Using the protein consensus sequence, degenerate nucleic acid probes are made to encode the protein consensus sequence, as is well known in the art. The protein sequence is encoded by DNA triplets which are deduced using standard tables. In some cases additional degeneracy is used to enable production in one oligonucleotide synthesis. In many cases motifs were chosen to minimize degeneracy. Amplification of neighboring sequences can utilize two motifs as indicated by faithful or error prone amplification. Alternatively outside sequences can be used as is indicated using vector sequence. In addition degenerate oligos can be synthesized and used directly in the procedure without amplification. Double stranded (ds) DNA probes are denatured and coated with RecA or another recombinase such as Rad5l. This material can be used to bind to and allow capture of specific clones from cDNA or genomic libraries. Alternatively this material can be introduced into cells producing transgenic cells or animals with alterations in related family members.

3.5.3.4.1.1.2. Second Subfamily Encoding Receptors That Bind Peptide Hormones That Do Not Show Sequence Similarity To The First R7G Subfamily

In addition to the first subfamily of G-protein coupled receptors, there is a second subfamily encoding receptors that bind peptide hormones that do not show sequence similarity to the first R7G subfamily. All the characterized receptors in this subfamily are coupled to G- proteins that activate both adenylyl cyclase and the phosphatidylinositol-calcium pathway. However, they are structurally similar; like classical R7G proteins they putatively contain seven transmembrane regions, a glycosylated extracellular N-terminus and a cytoplasmic C-terminus. Known receptors in this subfamily are encoded on multiple exons, and several of these genes are alternatively spliced to yield functionally distinct products. The N-terminus contains five conserved cysteine residues putatively important in disulfide bonds. Known G-protein coupled receptors in this subfamily are listed above.

3.5.3.4.1.1.3. Third Subfamily Encoding Receptors That Bind Glutamate And Calcium But Do Not Show Sequence Similarity To Either Of The Other Subfamilies

In addition to the first and second subfamilies of G-protein coupled receptors, there is a third subfamily encoding receptors that bind glutamate and calcium but do not show sequence similarity to either of the other subfamilies. Structurally, this subfamily has signal sequences, very large hydrophobic extracellular regions of about 540 to 600 amino acids that contain 17 conserved cysteines (putatively involved in disulfides), a region of about 250 residues that appear to contain seven transmembrane domains, and a C-terminal cytoplasmic domain of variable length (50 to 350 residues). Known G- protein coupled receptors of this subfamily are listed above.

3.5.3.4.1.2. Gene Family Is The bZIP Transcription Factor Family

In a preferred embodiment, the gene family is the bZIP transcription factor family. This eukaryotic gene family encodes DNA binding transcription factors that contain a basic region that mediates sequence specific DNA binding, and a leucine zipper, required for dimerization. The bZIP family includes, but is not limited to, AP-1, ATF, CREB, CREM, FOS, FRA, GBF, GCN4, HBP, JUN, MET4, OCS1, OP, TAF1, XBP1, and YBBO.

In a preferred embodiment, the gene family is involved in DNA mismatch repair, such as mutL, hexB and PMS1. Members of this family include, but are not limited to, MLH1, PMS1, PMS2, HexB and MulL. The protein consensus sequence is G-F-R-G-E-A-L.

In a preferred embodiment, the gene family is the mutS family, also involved in mismatch repair of DNA, directed to the correction of mismatched base pairs that have been missed by the proofreading element of the DNA polymerase complex. MutS gene family members include, but are not limited to, MSH2, MSH3, MSH6 and MutS. In a preferred embodiment, the gene family is the recA family. The bacterial recA is essential for homologous recombination and recombinatorial repair of DNA damage. RecA has many activities, including the formation of nucleoprotein filaments, binding to single stranded and double stranded DNA, binding and hydrolyzing ATP, recombinase activity and interaction with lexA causing lexA activation and autocatalytic cleavage. RecA family members include those from E.

coli, drosophila, human, lily, etc. specifically including but not limited to, E. coli recA, Recl, Rec2, Rad5l, Rad5l B, Rad5l C, Rad5l D, Rad5l E, XRCC2 and DMC1.

3.5.3.4.1.3. Gene Family Is The RecF Family

In a preferred embodiment, the gene family is the recF family. The prokaryotic recF protein is a single- stranded DNA binding protein which also putatively binds ATP. RecF is involved in DNA metabolism; it is required for recombinatorial DNA repair and for induction of the SOS response. RecF is a protein of about 350 to 370 amino acid residues; there is a conserved ATP-binding site motif "A" in the N-terminal section of the protein as well as two other conserved regions, one located in the central section and the other in the C-terminal section.

3.5.3.4.1.4. Gene Family Is The Bcl-2 Family

In a preferred embodiment, the gene family is the Bcl-2 family. Programmed cell death (PCD), or apoptosis, is induced by events such as growth factor withdrawal and toxins. It is generally controlled by regulators, which have either an inhibitory effect (i.e. anti-apoptotic) or block the protective effect of inhibitors (pro-apoptotic). Many viruses have found a way of countering defensive apoptosis by encoding their own anti-apoptotic genes thereby preventing their target cells from dying too soon.

All proteins belonging to the Bcl-2 family contain at least one of a BH1, BH2, BH3 or BH4 domain. All anti-apoptotic proteins contain BH1 and BH2 domains, some of them contain an additional N-terminal BH4 domain (such as Bcl-2, Bcl-x(L), Bcl-W, etc.), which is generally not found in pro-apoptotic proteins (with the exception of Bcl-x(S). Generally all pro-apoptotic proteins contain a BH3 domain (except for Bad), thought to be crucial for the dimerization of the proteins with other Bcl-2 family members and crucial for their killing activity. In addition, some of the pro- apoptotic proteins contain BH1 and BH2 domains (such as Bax and Bak). The BH3 domain is also present in some anti-apoptosis proteins, such as Bcl-2 and Bcl-x(L). Known Bcl-2 proteins include, but are not limited to, Bcl-2, Bcl-x(L), Bcl-W, Bcl- x(S), Bad, Bax, and Bak.

3.5.3.4.1.5. Gene Family Is The Site-Specific Recombinase Family

In a preferred embodiment, the gene family is the site-specific recombinase family. Site-specific recombination plays an important role in DNA rearrangement in prokaryotic organisms. Two types of site-specific recombination are known to occur: a) recombination between inverted repeats resulting in the reversal of a DNA segment; and b) recombination between repeat sequences on two DNA molecules resulting in their cointegration, or between repeats on one DNA molecule resulting the excision of a DNA fragment. Site-specific recombination is characterized by a strand exchange mechanism that requires no DNA synthesis or high energy cofactor; the phosphodiester bond energy is conserved in a phospho-protein linkage during strand cleavage and re- ligation.

Two unrelated families of recombinases are currently known. The first, called the "phage integrase" family, groups a number of bacterial, phage and yeast plasmid enzymes. The second, called the "resolvase" family, groups enzymes which share the following structural characteristics: an N-terminal catalytic and dimerization domain that contains a conserved serine residue involved in the transient covalent attachment to DNA, and a C-terminal helix-turn-helix DNA- binding domain.

3.5.3.4.1.6. Gene Family Is The Single-Stranded Binding Protein Family

In a preferred embodiment, the gene family is the single-stranded binding protein family. The E coli single-stranded binding protein (ssb), also known as the helix- destabilizing protein, is a protein of 177 amino acids. It binds tightly as a homotetramer to a single-stranded DNA ss-DNA) and plays an important role in DNA replication, recombination and repair. Members of the ssb family include, but are not limited to, E. coli ssb and eukaryotic RPA proteins.

3.5.3.4.1.7. Gene Family Is The TFIID Transcription Family

In a preferred embodiment, the gene family is the TFIID transcription family. Transcription factor TRID (or TATA-binding protein, TBP), is a general factor that plays a major role in the activation of eukaryotic genes transcribed by RNA polymerase II. TRID binds specifically to the TATA box promoter element which lies close to the position of transcription initiation. There is a remarkable degree of sequence conservation of a C-terminal domain of about 180 residues in TFIID from various eukaryotic sources. This region is necessary and sufficient for TATA box binding. The most significant structural feature of this domain is the presence of two conserved repeats of a 77 amino-acid region.

3.5.3.4.1.8. Gene Family Is The TGF-beta Family

In a preferred embodiment, the gene family is the TGF-beta family. Transforming growth factor-beta (TGF-beta) is a multifunctional protein that controls proliferation, differentiation and other functions in many cell types. TGF-beta-1 is a protein of 112 amino acid residues derived by proteolytic cleavage from the C-terminal portion of the precursor protein. Members of the TGF-beta family include, but are not limited to, the TGF-1-3 subfamily (including TGF1, TGF2, and TGF3); the BMP3 subfamily (BM3B, BMP3); the BMP5-8 subfamily (BM8A, BMP5, BMP6, BMP7, and BMP8); and the BMP 2 & 4 subfamily (BMP2, BMP4, DECA).

In a preferred embodiment, the gene family is the TNF family. A number of cytokines can be grouped into a family on the basis of amino acid sequence, as well as structural and functional similarities. These include (1) tumor necrosis factor (TNF), also known as cachectin or TNF-alpha, which is a cytokine with a wide variety of functions. TNF-alpha can cause cytolysis of certain tumor cell lines; it is involved in the induction of cachexia; it is a potent pyrogen, causing fever by direct action or by stimulation of interleukin-1 secretion; and it can stimulate cell proliferation and induce cell differentiation under certain conditions; (2) lymphotoxin-alpha (LT-alpha) and lymphotoxin-beta (LT-beta), two related cytokines produced by lymphocytes and which are cytotoxic for a wide range of tumor cells in vitro and in vivo; (3) T cell antigen gp39 (CD40L), a cytokine that seems to be important in B-cell development and activation; (4) CD27L, a cytokine that plays a role in T-cell activation; it induces the proliferation of costimulated T cells and enhances the generation of cytolytic T cells; (5) CD30L, a cytokine that induces proliferation of T- cells; (6) FASL, a cytokine involved in cell death; (8) 4-1 BBL, an inducible T cell surface molecule that contributes to T-cell stimulation; (9) OX40L, a cytokine that co- stimulates T cell proliferation and cytokine production; and (10), TNF-related apoptosis inducing ligand (TRAIL), a cytokine that induces apoptosis.

3.5.3.4.1.9. Gene Family Is The XPA Family

In a preferred embodiment, the gene family is the XPA family. Xeroderma pigmentosa (XP) is a human autosomal recessive disease, characterized by a high incidence of sunlight-induced skin cancer. Skin cells associated with this condition are hypersensitive to ultaviolet light, due to defects in the incision step of DNA excision repair. There are a minimum of 7 genetic complementation groups involved in this disorder: XPA to XPG. XPA is the most common form of the disease and is due to defects in a 30 kD nuclear protein called XPA or (XPAC). The sequence of XPA is conserved from higher eukaryotes to yeast (gene RAD14). XPA is a hydrophilic protein of 247 to 296 amino acid residues that has a C4- type zinc finger motif in its central section.

3.5.3.4.1.10. Gene Family Is The XPG Family

In a preferred embodiment, the gene family is the XPG family. The defect in XPG can be corrected by a 133 kD nuclear protein called XPG (or XPGC). Members of the XPG family include, but are not limited to, FEN1, XPG, RAD2, EXO1, and DIN7.

Once having identified a gene family and a consensus sequence, the compositions of the invention can be made. The compositions of the invention comprise at least one recombinase and at least two single-stranded targeting polynucleotides which are substantially complementary to each other and each have a consensus homology clamp for a gene family.

3.5.3.5. Homologous Recombination

Accordingly, the present invention provides methods of homologous recombination. By "homologous recombination" (HR) herein is meant an exchange of homologous or similar DNA sequence between two DNA molecules. An essential feature of HR is that the enzyme responsible for the recombination event can pair any homologous sequences as substrates. The ability of HR to transfer genetic information between DNA molecules makes targeted homologous recombination a very powerful method in genetic engineering and gene manipulation. HR can be used to insert, delete, and/or substitute any one or more nucleotides in a gene or gene segment or to introduce or delete genes in a targeted nucleic acid.

Once having identified a protein domain, the compositions of the invention can be made. The compositions of the invention comprise at least one recombinase and at least two single-stranded targeting polynucleotides which are substantially complementary to each other and each have a domain homology clamp.

3.5.3.6. Recombinase

By "recombinase" herein is meant a protein or peptide (e.g. L2 peptide) that, when included with an exogenous targeting polynucleotide, provide a measurable increase in the recombination frequency and/or localization frequency between the targeting polynucleotide and an endogenous predetermined DNA sequence. Thus, in a preferred embodiment, increases in recombination frequency from the normal range of 10^{-8} to 10^{-4}, to 10^{-4} to 10^{1}, preferably 10^{-3} to 10^{1}, and most preferably 10^{-2} to 10^{0}, may be achieved.

In the present invention, recombinase refers to a family of RecA-like and Rad5l -like recombination proteins all having essentially all or most of the same functions, particularly: (i) the recombinase protein's ability to properly bind to and position targeting polynucleotides on their homologous targets and (ii) the ability of recombinase protein/targeting polynucleotide complexes to efficiently find and bind to complementary endogenous sequences. The best characterized RecA protein is from E coli, in addition to the wild-type protein a number of mutant RecA proteins have been identified (e.g., RecA803; see Madiraju et al., PNAS USA 85(18):6592 (1988); Madiraju et al, Biochem. 31:10529 (1992); Lavery et al., J. Biol. Chem. 267:20648 (1992)). Further, many organisms have RecA-like recombinases with strand-transfer activities (e.g., Fugisawa et al., (1985) Nucl. Acids Res. 13: 7473; Hsieh et al., (1986) Cell 44: 885; Hsieh et al., (1989) J. Biol. Chem. 264: 5089; Fishel et al., (1988) Proc. Natl. Acad. Sci. (USA) 85: 3683; Cassuto et al., (1987) Mol. Gen. Genet. 208: 10; Ganea et al., (1987) Mol. Cell Biol. 7: 3124; Moore et al., (1990)J. Biol. Chem. 19:11108; Keene et al., (1984) Nucl. Acids Res. 12: 3057; Kimeic, (1984) Cold Spring Harbor Symp. 48: 675; Kmeic, (1986) Cell 44: 545; Kolodner et al., (1987) Proc. Natl. Acad. Sci. USA 84: 5560; Sugino et al., (1985) Proc. Natl. Acad. Sci. USA 85: 3683; Halbrook et al., (1989) J. Biol. Chem. 264: 21403; Eisen et al., (1988) Proc. Natl. Acad. Sci. USA 85: 7481; McCarthy et al., (1988) Proc. Natl. Acad. Sci. US 85: 5854; Lowenhaupt et al., (1989) J. Biol. Chem 264: 20568, which are incorporated herein by reference. Examples of such recombinase proteins include, for example but not limited to: RecA, RecA803, uvsX, and other RecA mutants and RecA-like recombinases (Roca, A. 1. (1990) Crit. Rev. Biochem. Molec. Biol. 25: 415), sep1 (Kolodner et al. (1987) Proc. NatI. Acad. Sci. (U.S.6.1 B4:5560; Tishkoff et al. Molec. Cell. Biol. 11:2593), RuvC (Dunderdale et al. (1991) Nature 354: 506), DST2, KEM1, XRN 1 (Dykstra et al. (1991) Molec. Cell. Biol. 11:2583), STPalpha/DST1 (Clark et al. (1991) Molec. Cell. Biol. 11:2576), HPP-1 (Moore et al.

(1991) Proc. Natl. Acad. Sci. (U.S.A.1 B8:9067), other target recombinases (Bishop et al. (1992) Cell 69 439; Shinohara et al. (1992) Cell 69 457); incorporated herein by reference. RecA may be purified from E coli strains, such as E coli strains JC 12772 and JC1 5369 (available from A.J. Clark and M. Madiraju, University of California-Berkeley, or purchased commercially). These strains contain the RecA coding sequences on a "runaway" replicating plasmid vector present at a high copy numbers per cell. The RecA803 protein is a high-activity mutant of wild- type RecA. The art teaches several examples of recombinase proteins, for example, from Drosophila, yeast, plant, human, and non-human mammalian cells, including proteins with biological properties similar to RecA (i.e., RecA-like recombinases), such as Rad51, Rad55, Rad57, dmcl from mammals and yeast. In addition, the recombinase may actually be a complex of proteins, i.e. a "recombinosome". In addition, included within the definition of a recombinase are portions or fragments of recombinases which retain recombinase biological activity, as well as variants or mutants of wild-type recombinases which retain biological activity, such as the E. coli RecA803 mutant with enhanced recombinase activity.

3.5.3.6.1. RecA or rad5l

In a preferred embodiment, RecA or rad5l is used. For example, RecA protein is typically obtained from bacterial strains that overproduce the protein: wild-type E coli RecA protein and mutant RecA803 protein may be purified from such strains. Alternatively, RecA protein can also be purchased from, for example, Pharmacia (Piscataway, NJ) or Boehringer Mannheim (Indianapolis, Indiana).

RecA proteins, and its homologs, form a nucleoprotein filament when it coats a single-stranded DNA. In this nucleoprotein filament, one monomer of RecA protein is bound to about 3 nucleotides. This property of RecA to coat single-stranded DNA is essentially sequence independent, although particular sequences favor initial loading of RecA onto a polynucleotide (e.g., nucleation sequences). The nucleoprotein filament(s) can be formed on essentially any DNA molecule and can be formed in cells (e.g., mammalian cells), forming complexes with both single- stranded and double-stranded DNA, although the loading conditions for dsDNA are somewhat different than for ssDNA.

3.5.3.6.1.1. The Recombinase Is Combined With Targeting Polynucleotides

The recombinase is combined with targeting polynucleotides as is more fully outlined below. By “nucleic acid" or"oligonucleotide" or"polynucleotide" or grammatical equivalents herein means at least two nucleotides covalently linked together. A nucleic acid of the present invention will generally contain phosphodiester bonds, although in some cases nucleic acid analogs are included that may have alternate backbones, comprising, for example, phosphoramide (Beaucage et al., Tetrahedron 49(10):1925 (1993) and references therein; Letsinger, J. Org. Chem. 35:3800 (1970); Sprinzl et al., Eur. J. Biochem. 81:579 (1977); Letsinger et al., Nucl. Acids Res. 14:3487 (1986); Sawai et al, Chem. Lett. 805 (1984), Letsinger et al., J. Am. Chem. Soc. 110:4470 (1988); and Pauwels et al., Chemica Scripta 26:141 91986)), phosphorothioate, phosphorodithioate, O- methylphophoroamidite linkages (see Eckstein, Oligonucleotides and Analogues: A Practical Approach, Oxford University Press), and peptide nucleic acid backbones and linkages (see Egholm, J. Am. Chem. Soc. 114:1895 (1992); Meier et al., Chem. Int. Ed. Engl. 31:1008 (1992); Nielsen, Nature, 365:566 (1993); Carlsson et al., Nature 380:207 (1996), all of which are incorporated by reference). These modifications of the ribose-phosphate backbone or bases may be done to facilitate the addition of other moieties such as chemical constituents, including 2' O-methyl and 5' modified substituents, as discussed below, or to increase the stability and half-life of such molecules in physiological environments.

The nucleic acids may be single stranded or double stranded, as specified, or contain portions of both double stranded or single stranded sequence. The nucleic acid may be DNA, both genomic and cDNA, RNA or a hybrid, where the nucleic acid contains any combination of deoxyribo-and ribo-nucleotides, and any combination of bases, including uracil, adenine, thymine, cytosine, guanine, inosine, xathanine and hypoxathanine, etc. Thus, for example, chimeric DNA-RNA molecules may be used such as described in Cole-Strauss et al., Science 273:1386 (1996) and Yoon et al., PNAS USA 93:2071 (1996), both of which are hereby incorporated by reference.

In general, the targeting polynucleotides may comprise any number of structures, as long as the changes do not substantially effect the functional ability of the targeting polynucleotide to result in homologous recombination. For example, recombinase coating of alternate structures should still be able to occur.

By "targeting polynucleotides" herein is meant the polynucleotides used to make alterations in the protein domains as described herein. Targeting polynucleotides are generally ssDNA or dsDNA, most preferably two complementary single-stranded DNAs.

Targeting polynucleotides are generally at least about 5 to 2000 nucleotides long, preferably about 12 to 200 nucleotides long, at least about 200 to 500 nucleotides long, more preferably at least about 500 to 2000 nucleotides long, or longer; however, as the length of a targeting polynucleotide increases beyond about 20,000 to 50,000 to 400,000 nucleotides, the efficiency or transferring an intact targeting polynucleotide into the cell decreases. The length of homology may be selected at the discretion of the practitioner on the basis of the sequence composition and complexity of the predetermined endogenous target DNA sequence(s) and guidance provided in the art, which generally indicates that 1.3 to 6.8 kilobase segments of homology are preferred when non- recombinase mediated methods are utilized (Hasty et al. (1991) Molec. Cell. Biol. 11: 5586; Shulman et al. (1990) Molec. Cell. Biol. 10: 4466, which are incorporated herein by reference).

Targeting polynucleotides have at least one sequence that substantially corresponds to, or is substantially complementary to, a predetermined endogenous DNA sequence. As used herein, the terms "predetermined target nucleic acid" and "predetermined target sequence" and "predetermined domain of a target nucleic acid" refer to polynucleotide sequences contained in a target nucleic acid. Such sequences include, for example, chromosomal sequences (e.g., structural genes, regulatory sequences including promoters and enhancers, recombinatorial hotspots, repeat sequences, integrated proviral sequences, hairpins, palindromes), episomal or extrachromosomal sequences (e.g., replicable plasmids or viral replication intermediates) including chloroplast and mitochondrial DNAsequences. By "predetermined" or"pre-selected" it is meant that the target sequence maybe selected at the discretion of the practitioner on the basis of known or predicted sequence information, and is not constrained to specific sites recognized by certain site- specific recombinases (e.g., FLIP recombinase or CRE recombinase). In some embodiments, the predetermined endogenous DNA target sequence will be other than a naturally occurring germline DNA sequence (e.g., a transgene, parasitic, mycoplasmal or viral sequence). An exogenous polynucleotide is a polynucleotide which is transferred into a target cell but which has not been replicated in that host cell; for example, a virus genome polynucleotide that enters a cell by fusion of a virion to the cell is an exogenous polynucleotide, however, replicated copies of the viral polynucleotide subsequently made in the infected cell are endogenous sequences (and may, for example, become integrated into a cell chromosome). Similarly, transgenes which are microinjected or transfected into a cell are exogenous polynucleotides, however integrated and replicated copies of the transgene(s) are endogenous sequences.

3.5.3.6.1.1.1. Target Nucleic Acid Comprises A Nucleotide Sequence Encoding A Protein Or Polypeptide Or Can Be Made To Comprise Non-Coding Regions As Well

In a preferred embodiment, the target nucleic acid comprises a nucleotide sequence encoding a protein or polypeptide, although as outlined herein, target nucleic acids may be made to non-coding regions as well. By "protein" herein is meant at least two covalently attached amino acids, which includes proteins, polypeptides, oligopeptides and peptides. Thus "amino acid" or "peptide residue", as used herein means naturally occurring and naturally modified amino acids. For example, "amino acid" also includes imino acid residues such as proline and hydroxyproline. A "naturally modified amino acid" includes for examples, amino acids that are modified to contain carbohydrate structures, such as high-mannose or complex carbohydrates, phosphate, or lipids. In the preferred embodiment, the amino acids are in the (S) or L-configuration.

The nucleotide sequence encoding the polypeptide is preferably operably linked to transcription and translation control elements operable in a host cell of interest, such that, introduction of the target nucleic acid results in expression of the encoded protein. The transcription control elements include a promoter, such as, a constitutive or inducible promoter. When the host cell of interest is a eukaryotic cell, enhancer elements are optionally employed. In a preferred embodiment the target nucleic acid is an extrachromosomal vector such as a plasmid. In other embodiments, the target nucleic acid is a viral vector, such as, a retrovirus, a phage, a BAC, PAC, YAC, MAC or other types of genomic and chromosomal DNA.

The term "naturally-occurring" as used herein as applied to an object refers to the fact that an object can be found in nature. For example, a polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a source in nature and which has not been intentionally modified by man in the laboratory is natural ly-occu rring.

3.5.3.6.1.1.2. The Target Nucleic Acid Comprises A Nucleic Acid Encoding A Protein Domain

The methods of the invention are used for alteration and evolution of protein domains; that is, in a preferred embodiment, the target nucleic acid comprises a nucleic acid encoding a protein domain. By "protein domain" and grammatical equivalents as used herein are meant a region of a protein that provides a specific structural and/or functional characteristic. Accordingly, a protein domain is an enzymatic active site, a ligand binding site, an allosteric effector region, an epitope, a region of a protein that is modified, such as, by addition of a carbohydrate, phosphate or lipid. A domain also relates to the hydrophobicity or hydrophilicity of a region and, therefore, also includes extracellular, intracellular, and transmembrane domains. Cell targeting sequences, such as, a signal peptide, nuclear localization sequence, mitochondrial localization sequences, etc. that direct proteins to either an extracellular or subcellular locale are domains. Additional domains include regions of proteins that interact with other proteins or nucleic acids, for example, include multimerization sequences, zinc- finger motifs, and the like. In another aspect, a protein domain is a region encoded by an exon.

Targeting polynucleotides have at least one sequence that substantially corresponds to, or is substantially complementary to, a target nucleic acid; in a preferred embodiment, it corresponds or complements a nucleic acid encoding a protein domain. By "corresponds to" herein is meant that a polynucleotide sequence is homologous (i.e., may be similar or identical, not strictly evolutionarily related) to all or a portion of a reference polynucleotide sequence, or that a polypeptide sequence is identical to a reference polypeptide sequence. In contradistinction, the term "complementary to" is used herein to mean that the complementary sequence can hybridize to all or a portion of a reference polynucleotide sequence. Thus, one of the complementary single stranded targeting polynucleotides is complementary to one strand of the endogenous target domain sequence (i.e. Watson) and corresponds to the other strand of the endogenous target domain sequence (i.e. Crick). Thus, the complementarity between two single-stranded targeting polynucleotides need not be perfect. For illustration, the nucleotide sequence "TATAC" corresponds to a reference sequence "TATAC" and is perfectly complementary to a reference sequence "GTATA.

The terms "substantially corresponds to" or "substantial identity" or "homologous" as used herein denotes a characteristic of a nucleic acid sequence, wherein a nucleic acid sequence has at least about 50 percent sequence identity as compared to a reference sequence, typically at least about 70 percent sequence identity, and preferably at least about 85 percent sequence identity as compared to a reference sequence. The percentage of sequence identity is calculated excluding small deletions or additions which total less than 25 percent of the reference sequence. The reference sequence may be a subset of a larger sequence, such as a portion of a gene or flanking sequence, or a repetitive portion of a chromosome. However, the reference sequence is at least 18 nucleotides long, typically at least about 30 nucleotides long, and preferably at least about 50 to 100 nucleotides long.

"Substantially complementary" as used herein refers to a sequence that is complementary to a sequence that substantially corresponds to a reference sequence. In general, targeting efficiency increases with the length of the targeting polynucleotide portion that is substantially complementary to a reference sequence present in the target DNA.

By "sequence homology" herein is meant sequence similarity or sequence identity.

Nucleic acid similarity can be determined using, for example, BLASTN (Altschul et aL 1990. J. Mol. Biol. 147:195-197). BLASTN uses a simple scoring system in which matches count +5 and mismatches -4. To achieve computational efficiency, the default parameters have been incorporated directly into the source code.

3.5.3.7. Percent Nucleic Acid Sequence Identity Is Determined

In an alternative embodiment, percent nucleic acid sequence identity is determined. In percent identity calculations relative weight is not assigned to the various types of sequence variation, such as, insertions, deletions, substitutions, etc. Only identities are scored positively (+1) and all forms of sequence variation given a value of "0", which obviates the need for a weighted scale or parameters as described above for sequence similarity calculations. Percent sequence identity can be calculated, for example, by dividing the number of matching identical residues by the total number of residues of the "shorter" sequence in the aligned region and multiplying by 100. The "longer" sequence is the one having the most actual residues in the aligned region.

3.5.3.8. Domain Homology Clamps: A Portion Of The Targeting Polynucleotide That Can Specifically Hybridize To A Nucleic Acid Encoding A Domain Within A Gene Of Interest

These corresponding/complementary sequences are sometimes referred to herein as "domain homology clamps", as they serve as templates for homologous pairing with the predetermined endogenous sequence(s). Thus, a "domain homology clamp" is a portion of the targeting polynucleotide that can specifically hybridize to a nucleic acid encoding a domain within a gene of interest. "Specific hybridization" is defined herein as the formation of hybrids between a targeting polynucleotide (e.g., a polynucleotide of the invention which may include substitutions, deletion, and/or additions as compared to the predetermined target nucleic acid sequence) and a predetermined target nucleic acid, wherein the targeting polynucleotide preferentially hybridizes to the predetermined target nucleic acid such that, for example, at least one discrete band can be identified on a Southern blot of nucleic acid prepared from target cells that contain the target nucleic acid sequence, and/or a targeting polynucleotide in an intact nucleus localizes to a discrete chromosomal location characteristic of a unique or repetitive sequence. As will be appreciated by those in the art, a target domain sequence may be present in more than one target polynucleotide species (e.g., a particular target sequence may occur in multiple members of a gene family). It is evident that optimal hybridization conditions will vary depending upon the sequence composition and length(s) of the targeting polynucleotide(s) and target(s), and the experimental method selected by the practitioner. Various guidelines may be used to select appropriate hybridization conditions (see, Maniatis et al., Molecular Cloning: A Laboratory Manual (1989), 2nd Ed., Cold Spring Harbor, N.Y. and Berger and Kimmel, Methods in Enzymology. Volume 152, Guide to Molecular Cloning Technigues (1987), Academic Press, Inc., San Diego, CA.), which are incorporated herein by reference. Methods for hybridizing a targeting polynucleotide to a discrete chromosomal location in intact nuclei are known in the art, see for example WO 93/05177 and Kowalczykowski and Zarling (1994) in Gene Targeting, Ed. Manuel Vega.

In targeting polynucleotides, domain homology clamps are typically located at or near the 5' or 3' end, preferably domain homology clamps are internal or located at each end of the polynucleotide (Berinstein et al. (1992) Molec, Cell. Biol. 12: 360, which is incorporated herein by reference). Without wishing to be bound by any particular theory, it is believed that the addition of recombinases permits efficient gene targeting with targeting polynucleotides having short (i. e., about 10 to 1000 basepair long) segments of homology, as well as with targeting polynucleotides having longer segments of homology.

3.5.3.9. Targeting Polynucleotides

3.5.3.9.1. Targeting Polynucleotides That Have Domain Homology Clamps That Are Highly Homologous To The Predetermined Target Endogenous Domain Functional Domain Nucleic Acid Sequence

Therefore, it is preferred that targeting polynucleotides of the invention have domain homology clamps that are highly homologous to the predetermined target endogenous domain functional domain nucleic acid sequence(s). Typically, targeting polynucleotides of the invention have at least one domain homology clamp that is at least about 18 to 35 nucleotides long, and it is preferable that domain homology clamps are at least about 20 to 100 nucleotides long, and more preferably at least about 100-500 nucleotides long, although the degree of sequence homology between the domain homology clamp and the targeted sequence and the base composition of the targeted sequence will determine the optimal and minimal clamp lengths (e.g., G-C rich sequences are typically more thermodynamically stable and will generally require shorter clamp length). Therefore, both domain homology clamp length and the degree of sequence homology can only be determined with reference to a particular predetermined sequence, but domain homology clamps generally must be at least about 10 nucleotides long and must also substantially correspond or be substantially complementary to a predetermined target sequence. Preferably, a homology clamp is at least about 10, and preferably at least about 50 nucleotides long and is substantially identical to or complementary to a predetermined target sequence. Without wishing to be bound by a particular theory, it is believed that the addition of recombinases to a targeting polynucleotide enhances the efficiency of homologous recombination between homologous, nonisogenic sequences (e.g., between an exon 2 sequence of an albumin gene of a Balb/c mouse and a homologous albumin gene exon 2 sequence of a C57/BL6 mouse), as well as between isogenic sequences.

3.5.3.9.2. Targeting Polynucleotides Comprising A Plurality Of Targeting Polynucleotides Comprising At Least One Shared Homology Clamp And A Degenerate Sequence

In one aspect of the invention, the targeting polynucleotides comprise a plurality of targeting polynucleotides comprising at least one shared homology clamp and a degenerate sequence. By "plurality" herein is meant more than one. The targeting polynucleotides find use in the mutagenesis and evolution of a target nucleic acid sequence that encodes specific protein domain by insertion, deletion and/or substitution of the nucleic acid sequence encoding the domain. In one embodiment the degenerate sequence is completely randomized, representing all possible combinations of nucleotides. In another embodiment, the degenerate sequence is biased, for example, to eliminate sequences encoding for transcriptional or translational stop signals. In another embodiment, the degenerate sequence is biased, to represent the codon bias of a host cell or class of organisms. The degenerate sequence is optionally biased to randomize specific sequence while maintaining other sequences constant. The length of the degenerate sequence is determined by the practitioner and is based on the desired number of nucleotides within the predetermined sequence to be modified.

3.5.3.9.3. Targeting Polynucleotides Substantially Identical To The Predetermined Target Sequence

In an alternative embodiment, the targeting polynucleotides are substantially identical to the predetermined target sequence. In the presence of a recombinase, the targeting polynucleotides form complexes with a predetermined target sequence of a target nucleic acid. As a part of the complex, the predetermined target sequence is resistant to nuclease digestion. The regions flanking the polynucleotide:target complex are susceptible to single-strand specific exonucleases. Accordingly, to effect domain specific evolution, these regions are nicked and the resultant fragments are reassembled and recombined by PCR as described below and by Stemmer et al. Nature. 370:389-391 and Stemmer et al. PNAS USA 91:10747-10751, hereby incorporated by reference.

The formation of heteroduplex joints is not a stringent process; genetic evidence supports the view that the classical phenomena of meiotic gene conversion and aberrant meiotic segregation results in part from the inclusion of mismatched base pairs in heteroduplex joints, and the subsequent correction of some of these mismatched base pairs before replication. Observations on RecA protein have provided information on parameters that affect the discrimination of relatedness from perfect or near-perfect homology and that affect the inclusion of mismatched base pairs in heteroduplex joints. The ability of RecA protein to drive strand exchange past all single base-pair mismatches and to form extensively mismatched joints in superhelical DNA reflect its role in recombination and gene conversion. This error-prone process may also be related to its role in mutagenesis. RecA-mediated pairing reactions involving DNA of X174 and G4, which are about 70 percent homologous, have yielded homologous recombinants (Cunningham et al. (1981) Cell 24: 213), although RecA preferentially forms homologous joints between highly homologous sequences, and is implicated as mediating a homology search process between an invading DNA strand and a recipient DNA strand, producing relatively stable heteroduplexes at regions of high homology. Accordingly, it is the fact that recombinases can drive the homologous recombination reaction between strands which are significantly, but not perfectly, homologous, which allows gene conversion and the modification of target sequences. Thus, targeting polynucleotides may be used to introduce nucleotide substitutions, insertions and deletions into an endogenous functional domain nucleic acid sequence, and thus the corresponding amino acid substitutions, insertions and deletions in proteins expressed from the endogenous domain functional domain nucleic acid sequence. By "endogenous" in this context herein is meant the naturally occurring sequence, i.e. sequences or substances originating from within a cell or organism. Similarly, "exogenous" refers to sequences or substances originating outside the cell or organism.

3.5.3.9.4. Method Where Two Substantially Complementary Targeting Polynucleotides Are Used

In a preferred embodiment, two substantially complementary targeting polynucleotides are used.

3.5.3.9.5. Method Where The Targeting Polynucleotides Form A Double Stranded Hybrid, Which May Be Coated With Recombinase

In one embodiment, the targeting polynucleotides form a double stranded hybrid, which may be coated with recombinase, although when the recombinase is

RecA, the loading conditions may be somewhat different from those used for single stranded nucleic acids.

3.5.3.9.6. Method Where Two Substantially Complementary Single- Stranded Targeting Polynucleotides Are Used

In a preferred embodiment, two substantially complementary single- stranded targeting polynucleotides are used. The two complementary single-stranded targeting polynucleotides are usually of equal length, although this is not required. However, as noted below, the stability of the four strand hybrids of the invention is putatively related, in part, to the lack of significant unhybridized single-stranded nucleic acid, and thus significant unpaired sequences are not preferred. Furthermore, as noted above, the complementarity between the two targeting polynucleotides need not be perfect. The two complementary single-stranded targeting polynucleotides are simultaneously or contemporaneously introduced into a target cell harboring a predetermined endogenous target sequence, generally with at lease one recombinase protein (e.g., RecA). Under most circumstances, it is preferred that the targeting polynucleotides are incubated with RecA or other recombinase prior to introduction into a target cell, so that the recombinase protein(s) may be "loaded" onto the targeting polynucleotide(s), to coat the nucleic acid, as is described below. Incubation conditions for such recombinase loading are described infra. A targeting polynucleotide may contain a sequence that enhances the loading process of a recombinase, for example a RecA loading sequence is the recombinogenic nucleation sequence poly[d(A-C)], and its complement, poly[d(G-T)]. The duplex sequence poly[d(A-C)*d(G-T)n, where n is from 5 to 25, is a middle repetitive element in target DNA.

There appears to be a fundamental difference in the stability of RecA- protein-mediated D-loops formed between one single-stranded DNA (ssDNA) probe hybridized to negatively supercoiled DNA targets in comparison to relaxed or linear duplex DNA targets. Internally located dsDNA target sequences on relaxed linear DNA targets hybridized by ssDNA probes produce single D- loops, which are unstable after removal of RecA protein (Adzuma, Genes Devel. 6:1679 (1992); Hsieh et al, PNAS USA 89:6492 (1992); Chiu et al., Biochemistry 32:13146 (1993)). This probe DNA instability of hybrids formed with linear duplex DNA targets is most probably due to the incoming ssDNA probe W-C base pairing with the complementary DNA strand of the duplex target and disrupting the base pairing in the other DNA strand. The required high free-energy of maintaining a disrupted DNA strand in an unpaired ssDNA conformation in a protein-free single-D-loop apparently can only be compensated for either by the stored free energy inherent in negatively supercoiled DNA targets or by base pairing initiated at the distal ends of the joint DNA molecule, allowing the exchanged strands to freely intertwine.

However, the addition of a second complementary ssDNA to the three- strand-containing single-D-loop stabilizes the deproteinized hybrid joint molecules by allowing W-C base pairing of the probe with the displaced target DNA strand. The addition of a second RecA-coated complementary ssDNA (cssDNA) strand to the three-strand containing single D-loop stabilizes deproteinized hybrid joints located away from the free ends of the duplex target DNA (Sena & Zarling, Nature Genetics 3:365 (1993); Revet et al. J. Mol. Biol. 232:779 (1993); Jayasena and Johnston, J. Mol. Bio. 230:1015 (1993)). The resulting four-stranded structure, named a double D-loop by analogy with the three-stranded single D-loop hybrid has been shown to be stable in the absence of RecA protein. This stability likely occurs because the restoration of W-C basepairing in the parental duplex would require disruption of two W-C basepairs in the double-D-loop (one W-C pair in each heteroduplex D-loop).

Since each base-pairing in the reverse transition (double-D-loop to duplex) is less favorable by the energy of one W-C basepair, the pair of cssDNA probes are thus kinetically trapped in duplex DNA targets in stable hybrid structures. The stability of the double-D loop joint molecule within internally located probe:target hybrids is an intermediate stage prior to the progression of the homologous recombination reaction to the strand exchange phase. The double D-loop permits isolation of stable multistranded DNA recombination intermediates.

In addition, when the targeting polynucleotides are used to generate insertions or deletions in an endogenous nucleic acid sequence, as is described herein, the use of two complementary single- stranded targeting polynucleotides allows the use of internal homology clamps. The use of internal homology clamps allows the formation of stable deproteinized cssDNA:probe target hybrids with homologous DNA sequences containing either relatively small or large insertions and deletions within a homologous DNA target. Without being bound by theory, it appears that these probe:target hybrids, with heterologous inserts in the cssDNA probe, are stabilized by the re-annealing of cssDNA probes to each other within the double-D-loop hybrid, forming a novel DNA structure with an internal homology clamp. Similarly stable double-D-loop hybrids formed at internal sites with heterologous inserts in the linear DNA targets (with respect to the cssDNA probe) are equally stable. Because cssDNA probes are kinetically trapped within the duplex target, the multi-stranded DNA intermediates of homologous DNA pairing are stabilized and strand exchange is facilitated.

3.5.3.10. Length Of The Internal Homology Clamp (i. e. The Length Of The Insertion Or Deletion)

In a preferred embodiment, the length of the internal homology clamp (i. e. the length of the insertion or deletion) is from about 1 to 50% of the total length of the targeting polynucleotide, with from about 1 to about 20% being preferred and from about 1 to about 10% being especially preferred, although in some cases the length of the deletion or insertion may be significantly larger. As for the domain homology clamps, the complementarity within the internal homology clamp need not be perfect. A targeting polynucleotide used in a method of the invention typically is a single-stranded nucleic acid, usually a DNA strand, or derived by denaturation of a duplex DNA, which is complementary to one (or both) strand(s) of the target duplex nucleic acid. Thus, one of the complementary single stranded targeting polynucleotides is complementary to one strand of the endogenous target sequence (i.e.Watson) and the other complementary single stranded targeting polynucleotide is complementary to the other strand of the endogenous target sequence (i.e. Crick). The domain homology clamp sequence preferably contains at least 90-95% sequence homology with the target sequence (although as outlined above, less sequence homology can be tolerated), to insure sequence-specific targeting of the targeting polynucleotide to the endogenous DNA domain target. Each single-stranded targeting polynucleotide is typically about 50-600 bases long, although a shorter or longer polynucleotide may also be employed.

3.5.3.11. Method For Making The Targeting Polynucleotides

Once the gene family and domain sequence is selected, the targeting polynucleotides are made, as will be appreciated by those in the art. For example, for large targeting polynucleotides, plasmids are engineered to contain an appropriately sized gene sequence with a deletion or insertion in the gene of interest and at least one flanking homology clamp which substantially corresponds or is substantially complementary to an endogenous target DNA sequence. Vectors containing a targeting polynucleotide sequence are typically grown in E coli and then isolated using standard molecular biology methods. Alternatively, targeting polynucleotides may be prepared in single-stranded form by oligonucleotide synthesis methods, which may first require, especially with larger targeting polynucleotides, formation of subfragments of the targeting polynucleotide, typically followed by splicing of the subfragments together, typically by enzymatic ligation. In general, as will be appreciated by those in the art, targeting polynucleotides may be produced by chemical synthesis of oligonucleotides, nick-translation of a double-stranded DNA template, polymerase chain-reaction amplification of a sequence (or ligase chain reaction amplification), purification of prokaryotic or target cloning vectors harboring a sequence of interest (e.g., a cloned cDNA or genomic clone, or portion thereof) such as plasmids, phagemids, YACs, cosmids, bacteriophage DNA, other viral DNA or replication intermediates, or purified restriction fragments thereof, as well as other sources of single and double-stranded polynucleotides having a desired nucleotide sequence. When using microinjection procedures it may be preferable to use a transfection technique with linearized sequences containing only modified target gene sequence and without vector or selectable sequences. The modified gene site is such that a homologous recombinant between the exogenous targeting polynucleotide and the endogenous DNA target sequence can be identified by using carefully chosen primers and PCR, followed by analysis to detect if PCR products specific to the desired targeted event are present (Erlich et al., (1991) Science 252: 1643, which is incorporated herein by reference). Several studies have already used PCR to successfully identify and then clone the desired transfected cell lines (Zimmer and Gruss, (1989) Nature 338: 150; Mouellic et al., (1990) Proc. Natl. Acad. Sci. USA 87: 4712; Shesely et al., (1991) Proc. Natl. Acad. Sci. USA 88: 4294, which are incorporated herein by reference). This approach is very effective when the number of cells receiving exogenous targeting polynucleotide(s) is high (i.e., with microinjection, or with liposomes) and the treated cell populations are allowed to expand to cell groups of approximately 1 X 10^4 cells (Capecchi, (1 1989) Science 244: 1288). When the target gene is not on a sex chromosome, or the cells are derived from a female, both alleles of a gene can be targeted by sequential inactivation (Mortensen et al., (1991) Proc. Natl. Acad. Sci. US 88: 7036). Alternatively, animals heterologous for the target gene can be bred to homologously as is known in the art.

The invention may also be practiced with individual targeting polynucleotides which do not comprise part of a complementary pair. In each case, a targeting polynucleotide is introduced into a target cell simultaneously or contemporaneously with a recombinase protein, typically in the form of a recombinase coated targeting polynucleotide as outlined herein (i.e., a polynucleotide pre-incubated with recombinase wherein the recombinase is noncovalently bound to the polynucleotide; generally referred to in the art as a nucleoprotein filament).

3.5.3.12. Alterations In The Target Nucleic Acid Comprising A Domain Or Domains Of Interest

The present invention allows for the introduction of alterations in the target nucleic acid comprising a domain or domains of interest. That is, the fact that heterologies are tolerated in targeting polynucleotides allows for two things: first, the use of a heterologous domain homology clamps that may target genes encoding functional domains of a protein or multiple proteins, resulting in a variety of genotypes and phenotypes, and secondly, the introduction of alterations to the target sequence. Thus typically, a targeting polynucleotide (or complementary polynucleotide pair) has a portion or region having a sequence that is not present in the preselected endogenous targeted sequence(s) (i.e., a nonhomologous portion or mismatch) which may be as small as a single mismatched nucleotide, several mismatches, or may span up to about several kilobases or more of nonhomologous sequence.

3.5.3.12.1. Methods And Compositions For Inactivation Of A Domain Of A Gene

Accordingly, in a preferred embodiment, the methods and compositions of the invention are used for inactivation of a domain of a gene. That is, exogenous targeting polynucleotides can be used to inactivate, decrease or alter the biological activity of one or more domains in a gene of a cell (or transgenic nonhuman animal or plant). This finds particular use in the generation of animal models of disease states, or in the elucidation of gene function and activity, similar to "knock out" experiments. Alternatively, the biological activity of the wild-type gene may be either decreased, or the wild-type activity altered to mimic disease states. This includes genetic manipulation of non-coding gene sequences that affect the transcription of genes, including, promoters, repressors, enhancers and transcriptional activating sequences.

3.5.3.12.1.1. Amino Acid Substitutions, Insertions Or Deletions In The Endogenous Target Sequences

Thus in a preferred embodiment, homologous recombination of the targeting polynucleotide and endogenous target sequence will result in amino acid substitutions, insertions or deletions in the endogenous target sequences, potentially both within the functional domain region and outside of it, for example as a result of the incorporation of PCR tags. This will generally result in modulated or altered gene function of the endogenous gene, including both a decrease or elimination of function as well as an enhancement of function. Nonhomologous portions are used to make insertions, deletions, and/or replacements in a predetermined endogenous targeted DNA sequence, and/or to make single or multiple nucleotide substitutions i n a predetermined endogenous target DNA sequence so that the resultant recombined sequence (i.e., a targeted recombinant endogenous sequence) incorporates some or all of the sequence information of the nonhomologous portion of the targeting polynucleotide(s). Thus, the nonhomologous regions are used to make variant sequences, i.e. targeted sequence modifications. In this way, site directed modifications may be done in a variety of systems for a variety of purposes.

3.5.3.12.1.1.1. Disruption By Either The Substitution, Insertion, Deletion Or Frame Shifting Of Nucleotides

The endogenous target sequence, generally nucleic acid encoding a domain, may be disrupted in a variety of ways. The term "disrupt" as used herein comprises a change in the coding or non-coding sequence of an endogenous nucleic acid. In one preferred embodiment, a disrupted gene will no longer produce a functional gene product. In another preferred embodiment, a disrupted gene produces a variant gene product. Generally, disruption may occur by either the substitution, insertion, deletion or frame shifting of nucleotides.

3.5.3.12.1.1.2. Disruption By Amino Acid Substitutions

In one embodiment, amino acid substitutions are made. This can be the result of either the incorporation of a non-naturally occurring domain sequence into a target, or of more specific changes to a particular sequence outside of the domain sequence.

3.5.3.12.1.1.3. Disruption By An Insertion Sequence

In one embodiment, the endogenous sequence is disrupted by an insertion sequence. The term "insertion sequence" as used herein means one or more nucleotides which are inserted into an endogenous gene to disrupt it. In general, insertion sequences can be as short as 1 nucleotide or as long as a gene, as outlined herein. For non-gene insertion sequences, the sequences are at least 1 nucleotide, with from about 1 to about 50 nucleotides being preferred, and from about 10 to 25 nucleotides being particularly preferred. An insertion sequence may comprise a polylinker sequence, with from about 1 to about 50 nucleotides being preferred, and from about 10 to 25 nucleotides being particularly preferred. Insertion sequence may be a PCR tag used for identification of the first gene. In a preferred embodiment, an insertion sequence comprises a gene which not only disrupts the endogenous gene, thus preventing its expression, but also can result in the expression of a new gene product. Thus, in a preferred embodiment, the disruption of an endogenous gene by an insertion sequence gene is done in such a manner to allow the transcription and translation of the insertion gene. An insertion sequence that encodes a gene may range from about 50 bp to 5000 bp of cDNA or about 5000 bp to 50000 bp of genomic DNA. As will be appreciated by those in the art, this can be done in a variety of ways. In a preferred embodiment, the insertion gene is targeted to the endogenous gene in such a manner as to utilize endogenous regulatory sequences, including promoters, enhancers or a regulatory sequence. In an alternate embodiment, the insertion sequence gene includes its own regulatory sequences, such as a promoter, enhancer or other regulatory sequence etc.

Particularly preferred insertion sequence genes include, but are not limited to, genes which encode selection or reporter proteins. In addition, the insertion sequence genes may be modified or variant genes.

3.5.3.12.1.1.4. Disruption By Deletions

The term "deletion" as used herein comprises removal of a portion of the nucleic acid sequence of an endogenous gene. Deletions range from about 1 to about

100 nucleotides, with from about 1 to 50 nucleotides being preferred and from about 1 to about 25 nucleotides being particularly preferred, although in some cases deletions may be much larger, and may effectively comprise the removal of the entire functional domain, the entire endogenous gene and/or its regulatory sequences. Deletions may occur in combination with substitutions or modifications to arrive at a final modified endogenous gene.

3.5.3.12.1.1.5. Disruption Simultaneously by An Insertion And A Deletion

In a preferred embodiment, endogenous genes may be disrupted simultaneously by an insertion and a deletion. For example, a domain of an endogenous gene, with or without its regulatory sequences, may be removed and replaced with an insertion sequence gene. Thus, for example, all but the regulatory sequences of an endogenous gene may be removed, and replaced with an insertion sequence gene, which is now under the control of the endogenous gene's regulatory elements.

The term "regulatory element" is used herein to describe a non-coding sequence which affects the transcription or translation of a gene including, but are not limited to, promoter sequences, ribosomal binding sites, transcriptional start and stop sequences, translational start and stop sequences, enhancer or activator sequences, dimerizing sequences, etc. In a preferred embodiment, the regulatory sequences include a promoter and transcriptional start and stop sequence. Promoter sequences encode either constitutive or inducible promoters. The promoters may be either naturally occurring promoters or hybrid promoters. Hybrid promoters, which combine elements of more than one promoter, are also known in the art, and are useful in the present invention.

In addition to domain homology clamps and optional internal homology clamps, the targeting polynucleotides of the invention may comprise additional components, such as cell-uptake components, chemical substituents, purification tags, etc.

3.5.3.12.2. Targeting Polynucleotide Comprising A Cell-Uptake Component

In a preferred embodiment, at least one of the targeting polynucleotides comprises at least one cell- uptake component. As used herein, the term "cell-uptake component" refers to an agent which, when bound, either directly or indirectly, to a targeting polynucleotide, enhances the intracellular uptake of the targeting polynucleotide into at least one cell type (e.g., hepatocytes). A targeting polynucleotide of the invention may optionally be conjugated, typically by covalently or preferably noncovalent , binding, to a cell-uptake component. Various methods have been described in the art for targeting DNA to specific cell types. A targeting polynucleotide of the invention can be conjugated to essentially any of several cell-uptake components known in the art. For targeting to hepatocytes, a targeting polynucleotide can be conjugated to an asialoorosomucoid (ASOR)-poly-L- lysine conjugate by methods described in the art and incorporated herein by reference (Wu GY and Wu CH (1987) J. Biol. Chem. 262:4429; Wu GY and Wu CH (1988) Biochemistry 27:887; Wu GY and Wu CH (1988) J. Biol. Chem. 263: 1462 1; Wu GY and Wu CH (1992) J. Biol. Chem. 267: 12436; Wu et al. (1991) J. Biol. Chem. 266: 14338; and Wilson et al. 0 992) J..Biol. Chem. 267: 963, W092/06180; W092/05250; and W091/17761, which are incorporated herein by reference).

Alternatively, a cell-uptake component may be formed by incubating the targeting polynucleotide with at least one lipid species and at least one protein species to form protein-lipid-polynucleotide complexes consisting essentially of the targeting polynucleotide and the lipid-protein cell-uptake component. Lipid vesicles made according to Feigner (W091/17424, incorporated herein by reference) and/or cationic lipidization (WO91/16024, incorporated herein by reference) or other forms for polynucleotide administration (EP 465,529, incorporated herein by reference) may also be employed as cell-uptake components. Nucleases, DNA damaging chemicals, UV radiation or gamma- radiation may also be used.

In addition to cell-uptake components, targeting components such as nuclear localization signals may be used, as is known in the art. See for example Kido et al., Exper. Cell Res. 198:107-114 (1992), hereby expressly incorporated by reference. Typically, a targeting polynucleotide of the invention is coated with at least one recombinase and is conjugated to a cell-uptake component, and the resulting cell targeting complex is contacted with a target cell under uptake conditions (e.g., physiological conditions) so that the targeting polynucleotide and the recombinase(s) are internalized in the target cell. A targeting polynucleotide may be contacted simultaneously or sequentially with a cell-uptake component and also with a recombinase; preferably the targeting polynucleotide is contacted first with a recombinase, or with a mixture comprising both a cell-uptake component and a recombinase under conditions whereby, on average, at least about one molecule of recombinase is noncovalently attached per targeting polynucleotide molecule and at least about one cell-uptake component also is noncovalently attached. Most preferably, coating of both recombinase and cell-uptake component saturates essentially all of the available binding sites on the targeting polynucleotide. A targeting polynucleotide may be preferentially coated with a cell-uptake component so that the resultant targeting complex comprises, on a molar basis, more cell-uptake component than recombinase(s). Alternatively, a targeting polynucleotide may be preferentially coated with recombinase(s) so that the resultant targeting complex comprises, on a molar basis, more recombinase(s) than cell- uptake component.

Cell-uptake components are included with recombinase-coated targeting polynucleotides of the invention to enhance the uptake of the recombinase-coated targeting polynucleotide(s) into cells, particularly for in vivo gene targeting applications, such as gene therapy to treat genetic diseases, including neoplasia, and targeted homologous recombination to treat viral infections wherein a viral sequence (e.g., an integrated hepatitis B virus (HBV) genome or genome fragment) may be targeted by homologous sequence targeting and inactivated. Alternatively, a targeting polynucleotide may be coated with the cell-uptake component and targeted to cells with a contemporaneous or simultaneous administration of a recombinase (e.g., liposomes or immunoliposomes containing a recombinase, a viral-based vector encoding and expressing a recombinase).

In addition to recombinase and cellular uptake components, at least one of the targeting polynucleotides may include chemical substituents. Exogenous targeting polynucleotides that have been modified with appended chemical substituents may be introduced along with recombinase (e.g., RecA) into a metabolically active target cell to homologously pair with a predetermined endogenous DNA target sequence in the cell. In a preferred embodiment, the exogenous targeting polynucleotides are derivatized, and additional chemical substituents are attached, either during or after polynucleotide synthesis, respectively, and are thus localized to a specific endogenous target sequence where they produce an alteration or chemical modification to a local DNA sequence. Preferred attached chemical substituents include, but are not limited to: cross-linking agents (see Podyminogin et al., Biochem. 34:13098 (1995) and 35:7267 (1996), both of which are hereby incorporated by reference), nucleic acid cleavage agents, metal chelates (e.g., iron/EDTA chelate for iron catalyzed cleavage), topoisomerases, endonucleases, exonucleases, ligases, phosphodiesterases, photodynamic porphyrins, chemotherapeutic drugs (e.g., adriamycin, doxirubicin), intercalating agents, labels, base-modification agents, agents which normally bind to nucleic acids such as labels, etc. (see for example Afonina et al., PNAS USA 93:3199 (1996), incorporated herein by reference) immunoglobulin chains, and oligonucleotides. Iron/EDTA chelates are particularly preferred chemical substituents where local cleavage of a DNA sequence is desired (Hertzberg et al. (1982) J. Am. Chem. Soc. 104: 313; Hertzberg and Dervan (1984) Biochemistry 23: 3934; Taylor et al. (1984) Tetrahedron 40: 457; Dervan, PB (1986) Science 232: 464, which are incorporated herein by reference). Further preferred are groups that prevent hybridization of the complementary single stranded nucleic acids to each other but not to unmodified nucleic acids; see for example Kutryavin et al., Biochem. 35:11170 (1996) and Woo et al., Nucleic Acid. Res. 24(13):2470 (1996), both of which are incorporated by reference. 2'-0 methyl groups are also preferred; see Cole-Strauss et al., Science 273:1386 (1996); Yoon et al., PNAS 93:2071 (1996)). Additional preferred chemical substituents include labeling moieties, including fluorescent labels. Preferred attachment chemistries include: direct linkage, e.g., via an appended reactive amino group (Corey and Schultz (1988) Science 238:1401, which is incorporated herein by reference) and other direct linkage chemistries, although streptavidin/biotin and digoxigenin/antidigoxigenin antibody linkage methods may also be used. Methods for linking chemical substituents are provided in U.S. Patents 5,135,720, 5,093,245, and 5, 055,556, which are incorporated herein by reference. Other linkage chemistries may be used at the discretion of the practitioner.

3.5.3.12.3. Targeting Polynucleotides Comprises At Least One Purification Tag Or Capture Moiety

In a preferred embodiment, at least one of the targeting polynucleotides comprises at least one purification tag or capture moiety, some of which are discussed above as chemical substituents, for example biotin, digoxigenin, psoralen, etc. Alternatively, the domain oligonucleotide could be directly attached to beads with the targeting reaction performed on a solid phase support.

3.5.3.12.4. Targeting Polynucleotides Are Coated With Recombinase Prior To Introduction To The Domain Target

In a preferred embodiment, the targeting polynucleotides are coated with recombinase prior to introduction to the domain target. The procedures below are directed to the use of E. coli RecA, although as will be appreciated by those in the art, other recombinases may be used as well. Targeting polynucleotides can be coated using GTPgammaS, mixes of ATPgammaS with rATP, rGTP and/or dATP, or dATP or rATP alone in the presence of an rATP generating system (Boehringer Mannheim). Various mixtures of GTPgammaS, ATPgammaS, ATP, AIDP, dATP and/or rATP or other nucleosides may be used, particularly preferred are mixes of ATPgammaS and ATP or ATPgammaS and ADP.

The targeting polynucleotide, whether double-stranded or sing le-stranded, is denatured by heating in an aqueous solution at 95- 100'C for five minutes, then placed in an ice bath for 20 seconds to about one minute followed by centrifugation at 0'C for approximately 20 sec, before use. When denatured targeting polynucleotides are not placed in a freezer at –20'C they are usually immediately added to standard RecA coating reaction buffer containing ATPgammaS, at room temperature, and to this is added the RecA protein.

Alternatively, RecA protein may be included with the buffer components and ATPgammaS before the polynucleotides are added.

RecA coating of targeting polynucleotide(s) is initiated by incubating polynucleotide-RecA mixtures at 37'Cfor 10-15min. RecA protein concentration tested during reaction with polynucleotide varies depending upon polynucleotide size and the amount of added polynucleotide, and the ratio of RecA molecule: nucleotide preferably ranges between about 3:1 and 1:3. When single-stranded polynucleotides are RecA coated independently of their homologous polynucleotide strands, the mM and microM concentrations of ATPgammaS and RecA, respectively, can be reduced to one-half those used with double-stranded targeting polynucleotides (i.e., RecA and ATPgammaS concentration ratios are usually kept constant at a specific concentration of individual polynucleotide strand, depending on whether a single- or double-stranded polynucleotide is used).

RecA protein coating of targeting polynucleotides is normally carried out in a standard 10x RecA coating reaction buffer. 10x RecA reaction buffer (i.e., 10x AC buffer) consists of. 100 mM Tris acetate (pH 7.5 at 37'C), 20 mM magnesium acetate, 500 mM sodium acetate, 10 mM DTT, and 50% glycerol). All of the targeting polynucleotides, whether double-stranded or single-stranded, typically are denatured before use by heating to 95-100'C for five minutes, placed on ice for one minute, and subjected to centrifugation (10,000 rpm) at 0'C for approximately 20 seconds (e.g., in a Tomy centrifuge). Denatured targeting polynucleotides usually are added immediately to room temperature RecA coating reaction buffer mixed with ATPgammaS and diluted with double- distilled H20 as necessary.

A reaction mixture typically contains the following components: (i) 0.2- 4.8 mM ATPgammaS; and (ii) between 1-100 ng/ul of targeting polynucleotide. To this mixture is added about 1-20,ul of RecA protein per 10-100 ul of reaction mixture, usually at about 2-10 mg/ml (purchased from Pharmacia or purified), and is rapidly added and mixed. The final reaction volume-for RecA coating of targeting polynucleotide is usually in the range of about 10-500 ul. RecA coating of targeting polynucleotide is usually initiated by incubating targeting polynucleotide-RecA mixtures at 37'C for about 10-15 min. RecA protein concentrations in coating reactions varies depending upon targeting polynucleotide size and the amount of added targeting polynucleotide: RecA protein concentrations are typically in the range of 5 to 50 uM. When single-stranded targeting polynucleotides are coated with RecA, independently of their complementary strands, the concentrations of ATPgammaS and RecA protein may optionally be reduced to about one-half of the concentrations used with double-stranded targeting polynucleotides of the same length: that is, the RecA protein and ATPgammaS concentration ratios are generally kept constant for a given concentration of individual polynucleotide strands.

3.5.3.12.4.1. Evaluation Of Coating Of Targeting Polynucleotides With RecA Protein

The coating of targeting polynucleotides with RecA protein can be evaluated in a number of ways. First, protein binding to DNA can be examined using band-shift gel assays (McEntee et al., (1981) 1. Biol. Chem. 256: 8835). Labeled polynucleotides can be coated with RecA protein in the presence of ATPgammaS and the products of the coating reactions may be separated by agarose gel electrophoresis.

Following incubation of RecA protein with denatured duplex DNAs the RecA protein effectively coats single-stranded targeting polynucleotides derived from denaturing a duplex DNA. As the ratio of RecA protein monomers to nucleotides in the targeting polynucleotide increases from 0, 1:27, 1:2.7 to 3.7:1 for 121-mer and 0, 1:22, 1:2.2 to 4.5:1 for 159-mer, targeting polynucleotide's electrophoretic mobility decreases, i.e., is retarded, due to RecA-binding to the targeting polynucleotide. Retardation of the coated polynucleotide's mobility reflects the saturation of targeting polynucleotide with RecA protein. An excess of RecA monomers to DNA nucleotides is required for efficient RecA coating of short targeting polynucleotides (Leahy et al., (1986) J. Biol. Chem. 261: 954).

A second method for evaluating protein binding to DNA is in the use of nitrocellulose fiber binding assays (Leahy et al., (1986) J. Biol. Chem. 261:6954; Woodbury, et al., (1983) Biochemistry 22(20):4730-4737. The nitrocellulose filter binding method is particularly useful in determining the dissociation-rates for protein:DNA complexes using labeled DNA. In the filter binding assay, DNA:protein complexes are retained on a filter while free DNA passes through the filter. This assay method is more quantitative for dissociation-rate determinations because the separation of DNA:protein complexes from free targeting polynucleotide is very rapid.

Alternatively, recombinase protein(s) (prokaryotic, eukaryotic or endogeneous to the target cell) may be exogenously induced or administered to a target cell simultaneously or contemporaneously (i.e., within about a few hours) with the targeting polynucleotide(s). Such administration is typically done by micro-injection, although electroporation, lipofection, and other transfection methods known in the art may also be used. Alternatively, recombinase-proteins may be produced in vivo. For example, they may be produced from a homologous or heterologous expression cassette in a transfected cell or targeted cell, such as a transgenic totipotent cell (e.g. a fertilized zygote) or an embryonal stem cell (e.g., a murine ES cell such as AB-1) used to generate a transgenic non- human animal line or a somatic cell or a pluripotent hematopoietic stem cell for reconstituting all or part of a particular stem cell population (e.g. hematopoietic) of an individual. Conveniently, a heterologous expression cassette includes a modulatable promoter, such as an ecdysone-inducible promoter- enhancer combination, an estrogen-induced promoter-enhancer combination, a CMV promoter- enhancer, an insulin gene promoter, or other cell-type specific, developmental stage-specific, hormone-inducible drug inducible, or other modulatable promoter construct so that expression of at least one species of recombinase protein from the cassette can by modulated for transiently producing recombinase(s) in vivo simultaneous or contemporaneous with introduction of a targeting polynucleotide into the cell. When a hormone-inducible promoter-enhancer combination is used, the cell must have the required hormone receptor present, either naturally or as a consequence of expression a co-transfected expression vector encoding such receptor. Alternatively, the recombinase may be endogeneous and produced in high levels. In this embodiment, preferably in eukaryotic target cells such as tumor cells, the target cells produce an elevated level of recombinase. In other embodiments the level of recombinase may be induced by DNA damaging agents, such as mitomycin C, UV or gamma- irradiation. Alternatively, recombinase, levels may be elevated by transfection of a plasmid encoding the recombinase gene into the cell.

3.5.3.13. Specialized Applications

3.5.3.13.1. Identification of New Members Of Gene Families Which May Be Useful In Functional Genomic Studies As Well As In The Identification Of New Drug Targets

Once made, the compositions of the invention find use in a number of applications upon administration to target cells. In general, the compositions and methods of the invention are useful to identify new members of gene families which may be useful in functional genomic studies as well as in the identification of new drug targets; both of these may be accomplished through the generation of "knock out" animal models. In addition, the present invention allows the modification of functional domain targets, the creation of transgenic plants and animals, the cloning of genes containing domain functional domains, etc.

3.5.3.13.2. Domain Specific Gene Evolution

Once made and administered to a target host cell, the compositions of the invention find use in a number of applications, including domain specific gene evolution. The polypeptide or protein encoded by the targeted nucleic undergoes homologous recombination with the plurality of polynucleotides to produce a plurality of modified target nucleic acids that are expressed to produce a plurality of modified proteins. Selection systems are employed to identify and isolate host cells expressing proteins having a desired property or phenotype. For example, if the expressed protein is an enzyme, cells having a modified enzyme activity are identified. The desired activity can be an increased or decreased or altered activity. Proteins having the desired phenotype are selected and isolated, the modified nucleic acid is sequenced to identify sequences effecting the desired activity, and the process is repeated iteratively as needed to produce a protein having a desired activity or property. In this and other embodiments, suitable target sequences include nucleic acid sequences encoding therapeutically or commercially relevant proteins, including, but not limited to, enzymes (proteases, recombinases, lipases, kinases, carbohydrases, isomerases, tautomerases, nucleases etc.), hormones, receptors, transcription factors, growth factors, cytokines, globin genes, immunosupppressive genes, tumor suppressors, oncogenes, complement- activating genes, milk proteins (casein, alpha-lactalbumin, beta-lactoglobulin, bovine and human serum albumin), immunoglobulins, milk proteins, and pharmaceutical proteins and vaccines.

In a preferred embodiment, the methods of the invention are used to generate pools or libraries of variant nucleic acid sequences, and cellular libraries containing the variant sequences. This idea is somewhat similar to the "gene shuffling" techniques of the literature (see Stemmer et al., 1994, Natuere 370:389 which attempt to rapidly "evolve" genes by making multiple random changes simultaneously. In the present invention, this end is accomplished by using at least one cycle, and preferably reiterative cycles, of enhanced homologous recombination with targeting polynucleotides containing random mismatches, substitutions, insertions, or deletions. By using a library of targeting polynucleotides comprising a plurality of random mutations, and repeating the homologous recombination steps as many times as needed, a rapid "gene evolution" can occur, wherein the new genes may contain large numbers of mutations.

Thus, in this embodiment, a plurality of targeting polynucleotides are used. The targeting polynucleotides each have at least one homology clamp that substantially corresponds to or is substantially complementary to the target sequence. Generally, the targeting polynucleotides are generated in pairs; that is, pairs of two single stranded targeting polynucleotides that are substantially complementary to each other are made (i.e. a Watson strand and a Crick strand). However, as will be appreciated by those in the art, less than a one to one ratio of Watson to Crick strands may be used; for example, an excess of one of the single stranded target polynucleotides (i.e. Watson) may be used. Preferably, sufficient numbers of each of Watson and Crick strands are used to allow the majority of the targeting polynucleotides to form double D-loops, which are preferred over single D- loops as outlined above. In addition, the pairs need not have perfect complementarity; for example, an excess of one of the single stranded target polynucleotides (i.e. Watson), which may or may not contain mismatches, may be paired to a large number of variant Crick strands, etc. Due to the random nature of the pairing, one or both of any particular pair of single- stranded targeting polynucleotides may not contain any mismatches. However, generally, at least one of the strands will contain at least one mismatch.

The plurality of pairs preferably comprise a pool or library of mismatches. The size of the library will depend on a number of factors, including the number of residues to be mutagenized, the succeptibility of the protein to mutation, etc., as will be appreciated by those in the art. Generally, a library in this instance preferably comprises at least 10% different mismatches over the length of the targeting polynucleotides, with at least 30% mismatches being preferred and at least 40% being particularly preferred, although as will be appreciated by those in the art, lower (1, 2, 5%, etc.) or higher amounts of mismatches being both possible and desirable in some instances. That is, the plurality of pairs comprise a pool of random and preferably degenerate mismatches over some regions or all of the entire targeting sequence. As outlined herein, "mismatches" include substitutions, insertions and deletions, with the former being preferred. Thus, for example, a pool of degenerate variant targeting polynucleotides covering some, or preferably all, possible mismatches over some region are generated, as outlined above, using techniques well known in the art. Preferably, but not required, the variant targeting polynucleotides each comprise only one or a few mismatches (less than 10), to allow complete multiple randomization. That is, by repeating the homologous recombination steps any number of times, as is more fully outlined below, the mismatches from a plurality of probes can be incorporated into a single target sequence.

The mismatches can be either non-random (i.e. targeted) or random, including biased randomness. That is, in some instances specific changes are desirable, and thus the sequence of the targeting polynucleotides are specifically chosen. In a preferred embodiment, the mismatches are random. The targeting polynucleotides can be chemically synthesized, and thus may incorporate any nucleotide at any position. The synthetic process can be designed to generate randomized nucleic acids, to allow the formation of all or most of the possible combinations over the length of the nucleic acid, thus forming a library of randomized targeting polynucleotides. Preferred methods maximize library size and diversity.

It is important to understand that in any library system encoded by oligonucleotide synthesis one cannot have complete control over the codons that will eventually be incorporated into the peptide structure. This is especially true in the case of codons encoding stop signals (TAA, TGA, TAG). In a synthesis with NNN as the random region, there is a 3/64, or 4.69%, chance that the codon will be a stop codon. To alleviate this, random residues are encoded as NNK, where K= T or G. This allows for encoding of all potential amino acids (changing their relative representation slightly), but importantly preventing the encoding of two stop residues TAA and TGA.

3.5.3.13.2.1. Mismatches Are Fully Randomized, With No Sequence Preferences Or Constants At Any Position

In one embodiment, the mismatches are fully randomized, with no sequence preferences or constants at any position.

3.5.3.13.2.2. Biased Library

In a preferred embodiment, the library is biased. That is, some positions within the sequence are either held constant, or are selected from a limited number of possibilities. For example, in a preferred embodiment, the nucleotides or amino acid residues are randomized within a defined class, for example, of hydrophobic amino acids, hydrophilic residues, sterically biased (either small or large) residues, towards the creation of cysteines, for cross-linking, prolines for SH-3 domains, serines, threonines, tyrosines or histidines for phosphorylation sites, etc., or to purines, etc.

As will be appreciated by those in the art, the introduction of a pool of variant targeting polynucleotides (in combination with recombinase) to a target sequence, in vitro to an extrachromosomal sequence, can result in a large number of homologous recombination reactions occuring over time. That is, any number of homologous recombination reactions can occur on a single target sequence, to generate a wide variety of single and multiple mismatches within a single target sequence, and a library of such variant target sequences, most of which will contain mismatches and be different from other members of the library. This thus works to generate a library of mismatches.

3.5.3.13.2.2.1. Generating A Large Number Of Different Variants Within A Particular Region Of A Sequence, Similar To Cassette Mutagenesis But Not Limited By Sequence Length

In a preferred embodiment, the variant targeting polynucleotides are made to a particular region or domain of a sequence (i.e. a nucleotide sequence that encodes a particular protein domain). For example, it may be desirable to generate a library of all possible variants of a binding domain of a protein, without affecting a different biologically functional domain, etc. Thus, the methods of the present invention find particular use in generating a large number of different variants within a particular region of a sequence, similar to cassette mutagenesis but not limited by sequence length. This idea is sometimes referred to herein as "domain specific gene evolution". In addition, two or more regions may also be altered simultaneously using these techniques; thus "single domain" and "multi- domain" shuffling can be done. Suitable domains include, but are not limited to, kinase domains, nucleotide-binding sites, DNA binding sites, signaling domains, receptor binding domains, transcriptional activating regions, promoters, origins, leader sequences, terminators, localization signal domains, and, in immunoglobulin genes, the complementarity determining regions (CDR), Fc, V_H, and V_L.

In a preferred embodiment, the variant targeting polynucleotides are made to the entire target sequence. In this way, a large number of single and multiple mismatches may be made in an entire sequence.

Thus, this embodiment proceeds as follows. A pool of , targeting polynucleotides are made each containing one or more mismatches. The probes are coated with recombinase as generally described herein, and introduced to the target sequence. Upon binding of the probes to form D-loops, the recombinase is preferably removed. These polynucleotide:target sequences can then introduced into recombinant proficient cells, to produce target protein which can then be tested for biological activity, based on the identification of the target sequence. Depending on the results, the altered target sequence can be used as the starting target sequence in reiterative rounds of homologous recombination, generally using the same library. Preferred embodiments utilize at least two rounds of homologous recombination, with at least 5 rounds being preferred and at least 10 rounds being particularly preferred. Again, the number of reiterative rounds that are performed will depend on the desired end-point, the resistance or succeptibility of the protein to mutation, the number of mismatches in each probe, etc.

3.5.3.14. Target Sequence

3.5.3.14.1. Target Sequences - An Immunoglobulin

In a preferred embodiment, the target sequence is an immunoglobulin. The amino terminal region of the light and heavy chains of an antibody that come together to form the antigen binding site and the variability of their amino acid sequences provides the structural basis for the diversity of antigen binding sites. The variability of the variable regions of both the heavy and light chains is for the most part restricted to three small hypervariable regions in each chain. The remianing part of the variable regions, known as framework regions, is relatively constant. Each of the hypervariable regions consists of only about 5 to 10 amino acids; the corresponding regions in the DNA encoding these regions are known as the complementarity determining regions, or CDRs. Thus to engineer an antibody library, for example an antibody phage library, one can change the sequences in the CDR regions of both the heavy and light chains. Different permutations and combinations of CDRs can be changed and evolved to engineer antibody-phage libraries.

3.5.3.14.2. Target Sequence - A Single-Chain Fv Framework For Any Number Of Specific Antigens

In a preferred embodiment, the target sequence is a single-chain Fv framework for any number of specific antigens. Single chain Fv (scFv) consists Of V_L and V_H domains of an immunoglobulin linked by a peptide spacer and thus contains the minimal antigen-binding domains of an antibody.

3.5.3.14.3. Target Sequence - An Antibody-Phage Fusion

In a preferred embodiment, antibody-phage fusions are used as the target sequence. As is known in the art, single-chain Fv fusions with the pill minor coat protein allows expression of the antibody on the surface of a phage, wherein it is available to bind antigen. Five copies of pill are expressed on the surface of the phage. It is therefor possible to express five scFv on the phage. This antibody-phage display system has been used previously to isolate novel antibodies. By starting with antibodies to any antigen, higher affinity antibodies may be made, as well as novel antibodies.

3.5.3.14.4. Target Sequence - The Coding Sequence For beta-Lactamase

In a preferred embodiment, the target sequence is the coding sequence for beta-lactamase.

Thus, the methods of the invention may be used to create superior recombinant reporter genes such as lacZ and green fluoroscent protein (GFP); superior antibiotic and drug resistance genes; superior recombinase genes; superior recombinant vectors; and other superior recombinant genes and proteins, including immunoglobulins, vaccines or other proteins with therapeutic value. For example, targeting polynucleotides containing any number of alterations may be made to one or more functional or structural domains of a protein, and then the products of homologous recombination evaluated.

Once made and administered to target cells, the target cells may be screened to identify a cell that contains the targeted sequence modification. This will be done in any number of ways, and will depend on the target gene and targeting polynucleotides as will be appreciated by those in the art. The screen may be based on phenotypic, biochemical, genotypic, or other functional changes, depending on the target sequence. In an additional embodiment, as will be appreciated by those in the art, selectable markers or marker sequences may be included in the targeting polynucleotides to facilitate later identification.

3.5.3.15. Kits Containing The Compositions Of The Invention Are Provided

In a preferred embodiment, kits containing the compositions of the invention are provided. The kits include the compositions, particularly those of libraries or pools of degenerate cssDNA probes, along with any number of reagents or buffers, including recombinases, buffers, ATP, etc.

3.5.3.16. Targeting Polynucleotide:Target Nucleic Acid Complexes Serve As Substrates For Single-Stranded Endonucleases

In an alternate embodiment, the targeting polynucleotide:target nucleic acid complexes serve as substrates for single-stranded endonucleases, such as, S1 and mung bean nuclease. Preferably the targeting polynucleotides are substantially complementary and form double D-loops with the target nucleic acid. The junctions of the complexes are single-stranded in nature, and thus are suceptible to single-strand specific nucleases and junction-specific nucleases. Accordingly, treatment of the complex with a single-strand nuclease results in defined nicks in the selected region encoding a predetermined domain of a protein encoded by the target nucleic acid. The nicked target nucleic acid is disassociated from the targeting polynucleotides and are reassembled and "shuffled" in vitro by PCR (Stemmer. 1994. Nature 370:389-391) to produce a plurality modified nucleic acids. The modified nucleic acids are introduced into an appropriate host cell, as described above, for expression of the plurality of modified proteins. Selection techniques are used as described herein to identify and isolate a cell expressing a modified protein. The process is repeated iteratively as needed to further evolve the targeted nucleic acid.

3.5.3.17. Isolation Of New Members Of Gene Families That Comprise Particular Domains

In a preferred embodiment, the present invention finds use in the isolation of new members of gene families that comprise particular domains. The use of domain filaments (i.e. domain homology clamps preferably containing a purification tag such as biotin, disoxisenin, or one purification method such as the use of a RecA antibody), allows the identification of genes containing the domain. Once identified, the new genes can be cloned, sequenced and the protein gene products purified. As will be appreciated by those in the art, the functional importance of the new genes can be assessed in a number of ways, including functional studies on the protein level, as well as the generation of "knock out" animal models. By choosing domain sequences for therapeutically relevant protein domains, novel targets can be identified that can be used in screening of drug candidates.

3.5.3.18. Utilizing The Purification Tag To Isolate The Gene(s)

Thus, in a preferred embodiment, the present invention provides methods for isolating new members of gene families containing protein domains comprising introducing targeting polynucleotides comprising domain homology clamps and at least one purification tag, preferably biotin, to a mix of nucleic acid, such as a plasmid cDNA library or a cell, and then utilizing the purification tag to isolate the gene(s). The exact methods will depend on the purification tag; a preferred method utilizes the attachment of the binding ligand for the tag to a bead, which is then used to pull out the sequence. Alternatively anti-RecA antibodies could be used to capture RecA-coated probes. The genes are then cloned, sequenced, and reassembled if necessary, as is well known in the art.

3.5.3.19. Use In Functional Genomic Studies, By Providing The Creation Of Transgenic Animal Models Of Disease

In an alternate preferred embodiment, the present invention finds use in functional genomic studies, by providing the creation of transgenic animal models of disease. Thus, for example, domain sequences used in homologous recombination methods can generate animals that have a wide variety of mutations in a wide variety of related domains of genes, potentially resulting in a wide variety of phenotypes, including phenotypes related to disease states. That is, by targeting a domain family, one, two or multiple genes in the family may be altered in any given experiment, thus creating a wide variety of genotypes and phenotypes to evaluate. Thus, in a preferred embodiment, the compositions and methods of the invention are used to generate pools or libraries of variant nucleic acid sequences, wherein the mutations are within the functional domain coding region, cellular libraries containing the variant libraries, and libraries of animals containing the variant libraries.

Furthermore, domain targeting can be used in cells or animals that are diseased or altered; in essence, domain targeting can be done to identify "reversion" genes, genes that can modulate disease states caused by domains of different genes. Thus for example the loss of one type of enzymatic activity, resulting in a disease phenotype, may be compensated by alterations in a different but homologous enzymatic activity. Accordingly, once the recombinase-targeting polynucleotide compositions are formulated, they are introduced or administered into target cells. The administration is typically done as is known for the administration of nucleic acids into cells, and, as those skilled in the art will appreciate, the methods may depend on the choice of the target cell. Suitable methods include, but are not limited to, microinjection, electroporation, lipofection, etc. By "target cells" herein is meant prokaryotic or eukaryotic cells. Suitable prokaryotic cells include, but are not limited to, bacteria such as E coli, Bacillus species, and the extremophile bacteria such as thermophiles, halophiles, etc. Preferably, the procaryotic target cells are recombination competent. Suitable eukaryotic cells include, but are not limited to, fungi such as yeast and filamentous fungi, including species of Aspergillus, Trichoderma, and Neurospora; plant cells including those of corn, sorghum, tobacco, canola, soybean, cotton, tomato, potato, alfalfa, sunflower, etc.; and animal cells, including fish, reptiles, amphibia, birds and mammals. Suitable fish cells include, but are not limited to, those from species of salmon, trout, tilapia, tuna, carp, flounder, halibut, swordfish, cod and zebrafish. Suitable bird cells include, but are not limited to, those of chickens, ducks, quail, pheasants, ostrich, and turkeys, and other jungle foul or game birds. Suitable mammalian cells include, but are not limited to, cells from horses, cows, buffalo, deer, sheep, rabbits, rodents such as mice, rats, hamsters and guinea pigs, goats, pigs, primates, marine mammals including dolphins and whales, as well as cell lines, such as human cell lines of any tissue or stem cell type, and stem cells, including pluripotent and non- pluripotent, and non-human zygotes. Particular human cells including, but are not limited to, tumor cells of all types (particularly melanoma, myeloid leukemia, carcinomas of the lung, breast, ovaries, colon, kidney, prostate, pancreas and testes), cardiomyocytes, endothelial cells, epithelial cells, lymphocytes (T-cell and B cell) , mast cells, eosinophils, vascular intimal cells, hepatocytes, leukocytes including mononuclear leukocytes, stem cells such as haemopoetic, neural, skin, lung, kidney, liver and myocyte stem cells, osteoclasts, chondrocytes and other connective tissue cells, keratinocytes, melanocytes, liver cells, kidney cells, and adipocytes. Suitable cells also include known research cells, including, but not limited to, Jurkat T cells, mouse La, HT1080, C127, Rat2, CV-1, NIH3T3 cells, CHO, COS, 293 cells, etc. See the ATCC cell line catalog, hereby expressly incorporated by reference.

3.5.3.20. Procaryotic Cells Are Used To Identify, Clone, Or Alter Target Sequences

In a preferred embodiment, procaryotic cells are used to identify, clone, or alter target sequences, preferably protein domains. In this embodiment, a pre-selected target DNA sequence is chosen for alteration. Preferably, the pre-selected target DNA sequence is contained within an extrachromosomal sequence. By "extrachromosomal sequence" herein is meant a sequence separate from the chromosomal or genomic sequences. Preferred extrachromosomal sequences include plasmids (particularly procaryotic plasmids such as bacterial plasmids), pl vectors, viral genomes, yeast, bacterial and mammalian artificial chromosomes (YAC, BAC and MAC, respectively), and other autonomously self-replicating sequences, although this is not required. As described herein, a recombinase and at least two single stranded targeting polynucleotides which are substantially complementary to each other, each of which contain a homology clamp to the target sequence contained on the extrachromosomal sequence, are added to the extrachromosomal sequence, preferably in vitro. The two single stranded targeting polynucleotides are preferably coated with recombinase, and at least one of the targeting polynucleotides contain at least one nucleotide substitution, insertion or deletion. The targeting polynucleotides then bind to the target sequence in the extrachromosomal sequence to effect homologous recombination and form an altered extrachromosomal sequence which contains the substitution, insertion or deletion. The altered extrachromosomal sequence is then introduced into the procaryotic cell using techniques known in the art. Preferably, the recombinase is removed prior to introduction into the target cell, using techniques known in the art. For example, the reaction may be treated with proteases such as proteinase K, detergents such as SDS, and phenol extraction (including phenol:chloroform:isoamyl alcohol extraction). These methods may also be used for eukaryotic cells. The cells are then grown under conditions which allow the expression of the variant nucleic acids to form variant proteins, particularly with alterations in domains.

3.5.3.20.1. Proteins Having The Desired Phenotype Are Selected And Isolated

In a preferred embodiment, proteins having the desired phenotype are selected and isolated, the modified nucleic acid is sequenced to identify sequences effecting the desired activity, and the process is repeated iteratively as needed to produce a protein having a desired activity or property. Thus, in a preferred embodiment, the methods of the invention are repeated until the desired protein or phenotype is seen.

Alternatively, the pre-selected target DNA sequence is a chromosomal sequence. In this embodiment, the recombinase with the targeting polynucleotides are introduced into the target cell, preferably eukaryotic target cells. In this embodiment, it may be desirable to bind (generally non-covalently) a nuclear localization signal to the targeting polynucleotides to facilitate localization of the complexes in the nucleus. See for example Kido et al., Exper. Cell Res. 198:107-114 (1992), hereby expressly incorporated by reference. The targeting polynucleotides and the recombinase function to effect homologous recombination, resulting in altered chromosomal or genomic sequences.

3.5.3.21. Eukaryotic Cells Are Used

In a preferred embodiment, eukaryotic cells are used. Basically, any mammalian cells may be used, with mouse, rat, primate and human cells being particularly preferred. Accordingly, suitable cell types include, but are not limited to, tumor cells of all types, i.e., fibroblasts, epithelial cells (particularly melanoma, myeloid leukemia, carcinomas of the lung, breast, ovaries, colon, kidney, prostate, pancreas and testes), cardiomyocytes, endothelial cells, epithelial cells, lymphocytes (T-cell and B cell) , mast cells, eosinophils, vascular intimal cells, hepatocytes, leukocytes including mononuclear leukocytes, stem cells such as haemopoetic, neural, skin, lung, kidney, liver and myocyte stem cells (for use in screening for differentiation and de-differentiation factors), osteoclasts, chondrocytes and other connective tissue cells, keratinocytes, melanocytes, liver cells, kidney cells, and adipocytes. Suitable cells also include known research cells, including, but not limited to, Jurkat T cells, NIH 3T3 cells, CHO, Cos, etc. See the ATCC cell line catalog, hereby expressly incorporated by reference.

For making transgenic non-human animals (which include homologously targeted non-human animals) embryonal stem cells (ES cells), donor cells for nuclear transfer and fertilized zygotes are preferred. In a preferred embodiment, embryonal stem cells are used. Murine ES cells, such as AB-1 line grown on mitotically inactive SNL76/7 cell feeder layers (McMahon and Bradley, Cell 62: 1073-1085 (1990)) essentially as described (Robertson, E.J. (1987) in Teratocarcinomas and Embcyonic Stem Cells: A Practical Approach. E.J. Robertson, ed. (oxford: IRL Press), p. 71-112; Zjilstra et al., Nature 342:435-438 (1989); and Schwartzberg et al., Science 246:799-803 (1989), each of which is incorporated herein by reference) may be used for homologous gene targeting. Other suitable ES lines include, but are not limited to, the E14 line (Hooper et al. (1987) Nature 326: 292-295), theD3 line (Doetschman et al. (1985) J. Embryol. Exp. Morph. 87: 21-45), and the CCE line (Robertson et al. (1986) Nature 323: 445-448). The success of generating a mouse line from ES cells bearing a specific targeted mutation depends on the pluripotence of the ES cells (i.e., their ability, once injected into a host blastocyst, to participate in embryogenesis and contribute to the germ cells of the resulting animal).

The pluripotence of any given ES cell line can vary with time in culture and the care with which it has been handled. The only definitive assay for pluripotence is to determine whether the specific population of ES cells to be used for targeting can give rise to chimeras capable of germline transmission of the ES genome. For this reason, prior to gene targeting, a portion of the parental population of AB-1 cells is injected into C57B1/6J blastocysts to ascertain whether the cells are capable of generating chimeric mice with extensive ES cell contribution and whether the majority of these chimeras can transmit the ES genome to progeny.

3.5.3.22. Non-Human Zygotes Are Used

In a preferred embodiment, non-human zygotes are used, for example to make transgenic animals, using techniques known in the art (see U.S. Patent No. 4,873,191; Brinster et al., PNAS 86:7007 (1989); Susulic et al., J. Biol. Chem. 49:29483 (1995), and Cavard et al. , Nucleic Acids Res. 16:2099 (1988), hereby incorporated by reference). Preferred zygotes include, but are not limited to, animal zygotes, including fish, avian, reptilian, amphibian and mammalian zygotes. Suitable fish zygotes include, but are not limited to, those from species of salmon, trout, tuna, carp, flounder, halibut, swordfish, cod, tilapia and zebrafish. Suitable bird zygotes include, but are not limited to, those of chickens, ducks, quail, pheasant, turkeys, and other jungle fowl and game birds. Suitable mammalian zygotes include, but are not limited to, cells from horses, cows, buffalo, deer, sheep, rabbits, rodents such as mice, rats, hamsters and guinea pigs, goats, pigs, primates, and marine mammals including dolphins and whales. See Hogan et al., Manipulating the Mouse Embryo (A Laboratory Manual), 2nd Ed. Cold Spring Harbor Press, 1994, incorporated by reference.

The vectors containing the DNA segments of interest can be transferred into the host cell by well-known methods, depending on the type of cellular host. For example, micro-injection is commonly utilized for target cells, although calcium phosphate treatment, electroporation, lipofection, biolistics or viral-based transfection also may be used. Other methods used to transform mammalian cells include the use of Polybrene, protoplast fusion, and others (see, generally, Sambrook et al. Molecular Cloning: A Laboratory Manual, 2d ed., 1989, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., which is incorporated herein by reference). Direct injection of DNA and/or recombinase-coated targeting polynucleotides into target cells, such as skeletal or muscle cells also may be used (Wolff et al. (1990) Science 247: 1465, which is incorporated herein by reference).

3.5.3.23. Precursor Animals Or Cells Already Contain A Disease Allele

In a preferred embodiment, the precursor animals or cells already contain a disease allele. As used herein, the term "disease allele" refers to an allele of a gene which is capable of producing a recognizable disease. A disease allele may be dominant or recessive and may produce disease directly or when present in combination with a specific genetic background or pre-existing pathological condition. A disease allele may be present in the gene pool or may be generated de novo in an individual by somatic mutation. For example and not limitation, disease alleles include: activated oncogenes, a sickle cell anemia allele, a Tay-Sachs allele, a cystic fibrosis allele, a Lesch-Nyhan allele, a retinoblastoma-susceptibility allele, a Fabry's disease allele, a Huntington's chorea allele, and a xenoderma pigmentosa allele. As used herein, a disease allele encompasses both alleles associated with human diseases and alleles associated with recognized veterinary diseases. For example, the deltaF508 CFTR allele in a human disease allele which is associated with cystic fibrosis in North Americans.

Once made and administered to target cells, new domains of genes may be isolated as outlined herein.

Alternatively, the target cells may be screened to identify a cell that contains the targeted functional domain sequence modification. This will be done in any number of ways, and will depend on the target domain and targeting polynucleotides as will be appreciated by those in the art. The screen may be based on phenotypic, biochemical, genotypic, or other functional changes, depending on the target sequence. For example, IgE levels may be evaluated for inflammation or asthma; vascular tone or blood pressure can be evaluated for hypertension, behavior screens can be done for neurologic effects, lipoprotein profiles can be screened for cardiovascular effects; secreted molecules can be evaluated for endocrine processes; CBCs can be done for hematology studies, etc. In an additional embodiment, as will be appreciated by those in the art, selectable markers or marker sequences may be included in the targeting polynucleotides to facilitate later identification.

The broad scope of this invention is best understood with reference to the following examples, which are not intended to limit the invention in any manner. All patents, patent applications, and publications cited herein are expressly incorporated by reference in their entirety.

3.5.4. Gene Deletion In Bacteria

This invention relates to a method and means for deleting a gene from a bacterial chromosome in a single step.

3.5.4.1 Applications

3.5.4.1.1 The Construction Of Special Bacterial Strains Which Have A Particular Genetic Background

Many applications require the construction of special bacterial strains which have a particular genetic background. These genetic backgrounds are the framework in which specific recombinant DNA plasmid constructions are tested to determine whether they can provide functions which are missing from the background of the bacteria. If such functions are provided by the recombinant plasmid, then there is positive evidence that a particular genetic locus or loci is encoded by the plasmid. The construction of genetic backgrounds is therefore a vital step in the subsequent cloning and investigation of specific genes.

3.5.4.1.2. The Gene In Question Has Been Deleted, So There Is No Production Whatsoever Of A Mutant Protein Making Analysis Much Less Ambiguous

The most common backgrounds are those in which a single mutation is present in a specific gene on the bacterial chromosome. This may result in synthesis of a defective version of the protein encoded by that gene, resulting in a specific cellular dysfunction. Correction of the cellular dysfunction, by introduction of a specific recombinant plasmid, is evidence that the relevant gene has been cloned onto the specific plasmid. However, a mutant protein may not be silent and may undergo interactions with other components, thereby creating the appearance that the plasmid gene encodes the entire active protein when it does not. This can seriously confuse the analysis. A much less ambiguous and therefore more desirable approach is one in which the gene in question has been deleted. In these circumstances, there is no production whatsoever of a mutant protein.

3.5.4.1.3. If Bacteria Is Being Used To Genetically Engineer A Protein, Deletion Of The Gene That Manufactures A Contaminating Protein Made By The Same Bacteria Can Reduce Purification Steps

Another situation where it is desirable to delete a complete gene from the chromosome of a bacteria is when the bacteria is being used in the production of a genetically engineered protein. Examples of these situations include the expression of insulin, growth hormone, protein A, and various vaccines from recombinant genes inserted into E. coli. Many times the E. coli produces proteins which contaminate the purified product produced by the genetic engineering. Although it is possible to add additional purification steps to remove this contaminant, it would be preferable to avoid the problem entirely by deleting the gene encoding the contaminating protein. Methods that are presently used to alter a gene include random mutation or inactivation of the gene sequence by mutation, insertion, or deletion of some portion of the gene. However, this can still lead to the production of inactive protein fragments or deletion of more of the chromosome than is necessary or desirable.

It is therefore an object of the present invention to provide a method and means for deleting a specific gene from a bacteria.

It is another object of the present invention to provide a method and means for inserting and/or inactivating or deleting the recA gene in a variety of bacteria.

3.5.4.2. Miscellaneous Applications

3.5.4.2.1 Creating A Deletion In A Defined Target Within The Bacterial Chromosome

There are two obstacles which have to be overcome. One is to develope a method which will create a deletion in a defined target within the bacterial chromosome.

3.5.4.2.2 Generalizing The Approach So That Even Essential Genes Can Be Deleted

The second is to generalize the approach so that even essential genes can be deleted. This is because many of the genes of interest are essential ones. The deletion of an essential gene normally would result in cell death.

Essential proteins includes enzymes of glycolysis, enzymes associated with amino acid or sugar biosynthesis, enzymes and factors associated with protein and nucleic acid biosynthesis (including both RNA and DNA), enzymes required for the synthesis of cofactors for oxidation, reduction, methylation and transamination processes, and enzymes necessary for synthesis of essential lipids and polysaccharides or of any other essential

molecule, including various nucleic acids, such as transfer or ribosomal RNAs, and segments of nucleic acids, such as gene regulatory elements.

It is therefore an object of the present invention to provide a method wherein an organism is produced which does not contain genetic material coding for the molecule which is to be cloned and expressed in the organism.

A further object of the present invention is to provide a method whereby a deficient organism which is to be used for cloning an essential gene remains viable even under restrictive conditions.

3.5.4.3 Specialized Applications

3.5.4.3.1. Method For The Deletion Of A Gene From A Bacteria Using A Single Step Procedure That Is Applicable To Any Gene That Has Been Cloned

Disclosed is a method for the deletion of a gene from a bacteria using a single step procedure that is applicable to any gene that has been cloned. The procedure depends upon site-directed recombination of linear DNA fragments with sequences on the chromosome as a function of recA in combination with the subsequent inactivation or deletion of the recA gene. The method is analogous to a procedure used to give insertions by homologous recombination into specific plasmid genes.

3.5.4.3.1.1. Strategy For Construction Of Chromosomal Deletions

The basic strategy for construction of chromosomal deletions is to transform the bacteria with linear DNA fragments which contain an antibiotic resistant or other phenotypically detectable gene segment (a "marker") flanked by sequences homologous to a closely spaced region on the cell chromosome containing the gene to be deleted A double-crossover event within the homologous sequences, effectively deleting the entire gene, is selected for by screening for the antibiotic resistant phenotype.

The linear fragment is not integrated into the chromosome in the absence of enzymes expressed by the recA gene. Accordingly, if the gene is absent or inactive, a recA gene must be inserted into the cell prior to the linear recombination event, and then inactivated or removed to prevent subsequent incorporation of other non-chromosomal sequences into the chromosome. This is particularly important if the bacteria is used as a host for the expression of genetically engineered proteins from sequences carried on plasmids or other extrachromosomal elements. The recA gene can be provided either in the form of an extrachromosomal element such as a plasmid or through incorporation of the gene into the chromosome. The recA gene is preferably inactivated or deleted by means of a double reciprocal recombination event utilizing linear sequences containing sequences homologous to the flanking sequences on either side of the recA gene in the chromosome. This is essentially the same method used to delete or insert a gene into the chromosome by homologous recombination as described above.

The present invention includes isolated linear DNA fragments constructed for use in the method for deleting a gene from the chromosome and for inserting or deleting the recA gene.

The method and sequences are applicable to a variety of bacteria including strains of Escherichia, Pseudomonas, Agrobacterium, Proteus, Erwinia, Shigella, Bacillus, Rhizobium, Vibrio, Salmonella, Streptococcus, and Haemophilus.
A plasmid which has a temperature-sensitive replicon and a wild-type allele of the desired gene is used to restore or maintain the phenotype produced by the deleted gene. This plasmid maintains production of the desired protein, and therefore cell viability if the encoded protein is essential to cell growth, when the chromosomal copy of the desired gene has been deleted. However, since the resulting cells have a temperature-sensitive phototype, the expression of the plasmid gene may be easily prevented by culturing the host strain at an elevated temperature. The resulting deficient host strain may then be used to screen other mutated and cloned genes for their ability to produce the desired protein.

3.5.5. Additional Considerations

3.5.5.1. Maintaining The Viability Of The Bacteria

The present invention is a method for deleting any gene from a bacterial strain, while maintaining the viability of the bacteria if the gene encodes an essential molecule and deletion of the essential gene results in a lethal phenotype. The gene to be deleted is provided on a plasmid with a temperature sensitive replicon. The cells now have a temperature sensitive phenotype. When the cells are grown at an elevated temperature under conditions which allow rapid detection of the absence of the desired molecule, complementation of this phenotype by introduction of a DNA fragment fused onto a stable plasmid is strong evidence for cloning of the gene which has been deleted from the chromosome.

3.5.5.2. Homologous Recombination

The present invention is a method for deleting any gene from a bacterial strain employing linear DNA fragments incorporating sequences homologous to the sequences flanking the gene to be deleted in the chromosome and sequences allowing insertion or removal/inactivation of recA in a variety of bacteria.

3.5.5.2.1. Roles For Several Enzymes Needed In Recombination

Homologous recombination has been detected in a wide variety of organisms, from simple bacteriophages to complex eukaryotic cells. Genetic and biochemical investigations have defined roles for several enzymes needed in recombination.

3.5.5.2.2. RecA

3.5.5.2.2.1. Alignment Of DNA Molecules Before Exchange

The RecA protein participates in the early steps of synapse, allowing alignment of DNA molecules before exchange, in strand transfer, where there is transfer of a single-stranded segment to a recipient duplex to form a limited heteroduplex region between the interacting DNAs and in the extension of this heteroduplex region by a reaction involving the concerted winding and unwinding of incoming and outgoing DNA chains, respectively. The hydrolysis of ATP by RecA protein is required for these events in vitro.

3.5.5.2.2.2. Controlling Expression Of A Group Of Unlinked Genes That Aid In Recovery Of Cells After Exposure To DNA Damaging Agents

The recA gene performs another equally important role in cell metabolism by controlling expression of a group of unlinked genes that aid in recovery of cells after exposure to DNA-damaging agents. This response, termed the SOS response, involves genes that participate in repair of DNA damage, mutagenesis, and coordination of cell division events.

3.5.5.2.2.3. Characterization

The purified RecA protein is a single polypeptide ranging in weight from about 37,000 to about 42,000. Although there is some variation in the sequences between bacterial strains, the recA proteins from a variety of bacteria in general have been isolated and characterized by interspecies complementation and assays utilizing comparisons with isolated, characterized proteins.

The present method and the linear fragments for use in the present method are not limited to organisms such as E. coli. The recA gene is found in a variety of bacteria including both gram negative and gram positive organisms. The recA gene has been isolated or identified and characterized in several organisms. It is possible to prepare a genomic library from any bacterial species and to isolate a clone containing a sequence homologous to a characterized recA gene by interspecific complementation using one of the available DNA clones containing a recA sequence or cross-reactivity with antisera to RecA proteins from a well-characterized organism such as E. coli.

RecA+ and RecA- strains are available from a variety of sources. For example, cloning and characterization of recA genes and recA proteins from Proteus vulgaris, Erwinia carotovoria, Shigella flexneria and Escherichia coli are described by S. L. Keener, et al., in J. Bacter. 160(1), 153-160 (1984). The RecA proteins produced by these organisms were demonstrated to be highly conserved among the species. In fact, the protein produced by one species could be introduced into another species where it complemented repair and regulatory defects of recA mutations. Other bacterial recA genes and gene products have been described by C.A. Miles, et al, in Mol. Gen.Genet. 204,161-165 (1986) (Agrobacterium tumefaciens C58), I. Goldberg et al, J. Bacteriol. 165(3), 715-722(1986), (Vibrio cholera), M. Better et al, J. Bacteriol. 155(1), 311-316 (1983), (Rhizobium meliloti), T.A. Kokjohn et al, J. Bacteriol. 163(2), 568-572 (1985), (Pseudomonas aeruginosa), and C.M. Lovett, Jr. et al, J. Bio.Chem. 260(6), 3305-3313 (1985) (Bacillus subtilis). These articles detail the isolation and characterization of gene libraries and the proteins encoded by the recA genes using techniques known to those skilled in the art including construction of gene libraries, identification of homologous genes using hybridization to probes from other more well characterized species such as E. coli, isolation and characterization of RecA proteins using antisera to RecA proteins from E. coli, and interspecies complementation of deficient strains of E. coli using gene segments from the libraries. The isolated proteins were useful for in vitro complementation studies. RecA deficient strains and RecA clones are available from many of the laboratories cited in the above articles and from the E. coli Genetic Stock Center at Yale University run by Dr. Barbara Bachman.

3.5.5.2.3. Construction Of Linear DNA Fragments Which Have Sequences Homologous To Closely Spaced Regions On The Chromosome In The Bacteria Which Flank Each Side Of The Gene Of Interest

The method whereby the DNA fragment is introduced into the chromosome to delete a gene or to regulate recA involves the construction of linear DNA fragments which have sequences homologous to closely spaced regions on the chromosome in the bacteria which flank each side of the gene of interest. In general, they must contain at least 80 to 100 nucleotides homologous to sequences flanking the gene to be deleted or the recA gene. An antibiotic resistant locus such as Kan^r, which encodes a protein making the bacterial resistant to the antibiotic, or other marker, is placed between these flanking sequences.

The linear DNA segment is introduced into a specific cell strain and selection is done for a double, reciprocal recombination which will delete the target gene and insert the marker into its place. As a result of the insertion of the antibiotic resistant gene, the cells are now resistant to the antibiotic. Cells which do not contain the insertion are eliminated by growing the bacteria in a medium containing the antibiotic.

Any other gene which allows for rapid screening of the cells containing a double, reciprocal recombination may be used in place of a gene for antibiotic resistance. For example, any gene for an essential protein, including enzymes, cofactors, proteins which are necessary for the synthesis of essential lipids, polysaccharides, nucleic acids, and other protein molecules such as receptors, as well as nucleic acids which have functional activity such as ribozymes, may be used. Other genes which confer a detectable phenotype on the cell strain such as sensitivity to temperature or ultraviolet radiation, auxotrophism for a sugar, amino acid, protein or nucleotide, or any other phenotype which can be detected by chemical indicators either in vitro or in vivo assay, or an immunoassay for a specific cellular component may also be used. Such chemical, radioactive, or immunological screening assays are well known to those skilled in the art.

3.5.5.3. Eliminating Contamination From The Host Producing Its Own Protein By Deletion And Replacement

In one application, in which the goal is to produce and purify a foreign protein, and the microorganism encodes its own version of the protein, the gene for the microorganism's own protein is eliminated. The gene for the foreign protein is then inserted and the protein produced. The purification process is thereby simplified since there is no contamination by the host protein, whether analogous to the protein being produced or unrelated which copurifies with, or interferes with the purification of, the protein being produced. For example, a protein may interfere with binding of the protein to be purified to a column.

3.5.5.4. A Plasmid Is Used To Introduce A Mutated Gene Into An Organism

In a second application, a plasmid is used to introduce a gene into an organism which typically contains a mutation in the gene to be investigated resulting in a negative phenotype for the product of the gene to be investigated. Failure of the organism to produce a biologically active form of the protein encoded by the mutated gene may confer a lethal phenotype under certain defined conditions. For example, this can be at a temperature, designated as the restrictive temperature, at which the mutant protein denatures or otherwise undergoes inactivation. The cloned gene which is introduced is selected for by virtue of its ability to confer cell viability or any other detectable phenotype for the desired protein at the restrictive temperature. The acquisition of viability

or other detectable phenotype at the normally restrictive temperature is evidence that the gene of interest has been cloned.

3.5.5.4.1. Problems: The Defective Host Protein May Interact With A Protein Produced From The Introduced Plasmid

The major problem with this second system is that, unless the inactive gene is deleted in entirety, the defective host protein may interact with a protein produced from the introduced plasmid. This interaction may stabilize the defective host protein enough so that its activity is restored even at the restrictive temperature. In this case, the restoration of growth at the restrictive temperature would be a false positive, that is, the growth would not be due to activity encoded by the cloned DNA segment. This problem holds true for all selections based on complementation of a phenotype which is due to a defect in a specific protein. Such phenotypes include temperature sensitivity, amino acid auxotrophies or any other auxotrophies which result from a lack of synthesis of a key ingredient such as a sugar, nucleotide, critical protein or nucleic acid, or cofactor used for oxidation, reduction, or transamination reactions.

3.5.5.5. False Positives

The problem of "false positives" also exists for cloned DNA pieces which are created to encode enzyme fragments as a means to define the catalytic core or to define any segment which achieves a specific purpose, such as a piece which undergoes self-association, binds to a specific ligand or receptor, or forms a specific complex or array with one or more additional components. In these cases, the engineering of protein fragments, which are tested in a host cell that encodes a defective version of the protein of interest, is seriously hampered if the defective host protein interacts in any way with the engineered pieces.

3.5.5.6. Use Of Temperature Sensitive Replicons

In the present invention, specifically designed linear DNA fragments are used to create a deletion of a gene by site-specific recombination. These fragments are transformed into the host cell. Cell viability or the detectable phenotype can be maintained during the procedure by provision of the gene encoding the desired protein on a recombinant plasmid that has a temperature-sensitive replicon, so that the cells which contain the deletion have a temperature sensitive phenotype. To achieve the deletion by recombination with the linear DNA fragments, it is necessary for the cells to have a RecA+ phenotype which is derived from recA, or its equivalent. Once recombination has occurred, the cell must immediately be changed to RecA or else the temperature sensitive plasmid will recombine with homologous sequences on the chromosome. The same would apply to any other extrachromosomal element where integration into the host chromosome would be undesirable.

The RecA- phenotype may be achieved by simultaneous inactivation of recA during the transformation with linear fragments or, after the transformation, by immediately introducing RecA- by mating with an appropriate RecA- strain or by transduction with a phage which carries a RecA- gene segment. Although mutagenesis may also be an effective means of making the cell RecA-, this is a "hit or miss" approach. The preferred method is to use homologous recombination of linear DNA sequences bounded by sequences hybridizing to the sequences flanking the recA gene. The recA gene is necessary in order for the gene encoding the desired protein to be incorporated into the organism. However, any plasmids or other extrachromosomal elements in the cell will be incorporated unless the recA gene is immediately removed. This is a particular concern where the bacteria serves as a host for the expression of a genetically engineered protein from multicopy plasmids.

3.6. Artificial Chromosomes

3.6.1 Artificial chromosomes, uses thereof and methods for preparing artificial chromosomes

Methods for preparing cell lines that contain artificial chromosomes, methods for preparation of artificial chromosomes, methods for isolation of the artificial chromosomes, methods for purification of artificial chromosomes, methods for targeted insertion of heterologous DNA into artificial chromosomes, methods for delivery of the chromosomes to selected cells and tissues, and methods for isolation and large-scale production of the chromosomes are provided. Also provided are cell lines for use in the methods, and cell lines and chromosomes produced by the methods. In particular, satellite artificial chromosomes that, except for inserted heterologous DNA, are substantially composed of heterochromatin are provided. Cell-based methods for use of the artificial chromosomes, including for gene therapy, production of gene products and production of transgenic plants and animals are also provided.

3.6.2 Limitations Of Existing Gene Delivery Technologies

Several viral vectors, non-viral, and physical delivery systems for gene therapy and recombinant expression of heterologous nucleic acids have been developed [see, e.g., Mitani et al. (1993) Trends Biotech. 11:162-166]. The presently available systems, however, have numerous limitations, particularly where persistent, stable, or controlled gene expression is required. These limitations include: (1) size limitations because there is a limit, generally on order of about ten kilobases [kB], at most, to the size of the DNA insert [gene] that can be accepted by viral vectors, whereas a number of mammalian genes of possible therapeutic importance are well above this limit, especially if all control elements are included; (2) the inability to specifically target integration so that random integration occurs which carries a risk of disrupting vital genes or cancer suppressor genes; (3) the expression of randomly integrated therapeutic genes may be affected by the functional compartmentalization n the nucleus and are affected by chromatin-based position effects; (4) the copy number and consequently the expression of a given gene to be integrated into the genome cannot be controlled. Thus, improvements in gene delivery and stable expression systems are needed [see, e.g., Mulligan (1993) Science 260:926-932].

In addition, safe and effective vectors and gene therapy methods should have numerous features that are not assured by the presently available systems. For example, a safe vector should not contain DNA elements that can promote unwanted changes by recombination or mutation in the host genetic material, should not have the potential to initiate deleterious effects in cells, tissues, or organisms carrying the vector, and should not interfere with genomic functions. In addition, it would be advantageous for the vector to be non-integrative, or designed for site-specific integration. Also, the copy number of therapeutic gene(s) carried by the vector should be controlled and stable, the vector should secure the independent and controlled function of the introduced gene(s); and the vector should accept large (up to Mb size) inserts and ensure the functional stability of the insert.

The limitations of existing gene delivery technologies, however, argue for the development of alternative vector systems suitable for transferring large [up to Mb size or larger] genes and gene complexes together with regulatory elements that will provide a safe, controlled, and persistent expression of the therapeutic genetic material.

At the present time, none of the available vectors fulfill all these requirements. Most of these characteristics, however, are possessed by chromosomes. Thus, an artificial chromosome would be an ideal vector for gene therapy, as well as for stable, high-level, controlled production of gene products that require coordination of expression of numerous genes or that are encoded by large genes, and other uses. Artificial chromosomes for expression of heterologous genes in yeast are available, but construction of defined mammalian artificial chromosomes has not been achieved. Such construction has been hindered by the lack of an isolated, functional, mammalian centromere and uncertainty regarding the requisites for its production and stable replication. Unlike in yeast, there are no selectable genes in close proximity to a mammalian centromere, and the presence of long runs of highly repetitive pericentric heterochromatic DNA makes the isolation of a mammalian centromere using presently available methods, such as chromosome walking, virtually impossible. Other strategies are required for production of mammalian artificial chromosomes, and some have been developed. For example, U.S. Pat. No. 5,288,625 provides a cell line that contains an artificial chromosome, a minichromosome, that is about 20 to 30 megabases. Methods provided for isolation of these chromosomes, however, provide preparations of only about 10-20% purity. Thus, development of alternative artificial chromosomes and perfection of isolation and purification methods as well as development of more versatile chromosomes and further characterization of the minichromosomes is required to realize the potential of this technology.

Therefore, it is an object herein to provide mammalian artificial chromosomes and methods for introduction of foreign DNA into such chromosomes. It is also an object herein to provide methods of isolation and purification of the chromosomes. It is also an object herein to provide methods for introduction of the mammalian artificial chromosome into selected cells, and to provide the resulting cells, as well as transgenic animals, birds, fish and plants that contain the artificial chromosomes. It is also an object herein to provide methods for gene therapy and expression of gene products using artificial chromosomes. It is a further object herein to provide methods for constructing species-specific artificial chromosomes de novo. Another object herein is to provide methods to generate de novo mammalian artificial chromosomes.

3.6.3Methods For Preparing Artificial Chromosomes

Mammalian artificial chromosomes [MACs] are provided. Also provided are artificial chromosomes for other higher eukaryotic species, such as insects, birds, fowl and fish, produced using the MACS and methods provided herein. Methods for generating and isolating such chromosomes are provided.

Methods using the MACs to construct artificial chromosomes from other species, such as insect, bird, fowl and fish species are also provided. The artificial chromosomes are fully functional stable chromosomes. Two types of artificial chromosomes are provided. One type, herein referred to as SATACs [satellite artificial chromosomes] are stable heterochromatic chromosomes, and the other type are minichromosomes based on amplification of euchromatin.

Artificial chromosomes provide an extra-genomic locus for targeted integration of megabase pair size DNA fragments that contain single or multiple genes, including multiple copies of a single gene operatively linked to one promoter or each copy or several copies linked to separate promoters. Thus, methods using the MACs to introduce the genes into cells, tissues, and animals, as well as species such as birds, fowl, fish and plants, are also provided. The artificial chromosomes with integrated heterologous DNA may be used in methods of gene therapy, in methods of production of gene products, particularly products that require expression of multigenic biosynthetic pathways, and also are intended for delivery into the nuclei of germline cells, such as embryo-derived stem cells [ES cells], for production of transgenic animals, birds, fowl and fish. Transgenic plants, including monocots and dicots, are also contemplated herein.

Mammalian artificial chromosomes provide extra-genomic specific integration sites for introduction of genes encoding proteins of interest and permit megabase size DNA integration so that, for example, genes encoding an entire metabolic pathway or a very large gene, such as the cystic fibrosis [CF; about 250 kb] genomic DNA gene [cystic fibrosis [CF; about 600 kb] gene], several genes, such as multiple genes encoding a series of antigens for preparation of a multivalent vaccine, can be stably introduced into a cell. Vectors for targeted introduction of such genes, including the tumor suppressor genes, such as p53, the cystic fibrosis transmembrane regulator cDNA [CFTR], and the genes for anti-HIV ribozymes, such as an anti-HIV gag ribozyme gene, into the artificial chromosomes are also provided.

The chromosomes provided herein are generated by introducing heterologous DNA that includes DNA encoding one or multiple selectable marker(s) into cells, preferably a stable cell line, growing the cells under selective conditions, and identifying from among the resulting clones those that include chromosomes with more than one centromere and/or fragments thereof. The amplification that produces the additional centromere occurs in cells that contain chromosomes in which the heterologous DNA has integrated near the centromere in the pericentric region of the chromosome. The selected clonal cells are then used to generate artificial chromosomes.

In preferred embodiments, the DNA with the selectable marker that is introduced into cells to generate artificial chromosomes includes sequences that target it to the pericentric region of the chromosome. For example, vectors, such as pTEMPUD, which includes such DNA specific for mouse satellite DNA, are provided. Also provided are derivatives of pTEMPUD containing human satellite DNA sequences that specifically target human chromosomes or human satellite sequences. Upon integration, these vectors can induce the amplification that results in generation of additional centromeres.

Artificial chromosomes are generated by culturing the cells with the multicentric, typically dicentric, chromosomes under conditions whereby the chromosome breaks to form a minichromosome and formerly dicentric chromosome. Among the MACs provided herein are the SATACs, which are primarily made up of repeating units of short satellite DNA and are fully heterochromatic, so that without insertion of heterologous or foreign DNA, the chromosomes preferably contain no genetic information. They can thus be used as "safe" vectors for delivery of DNA to mammalian hosts because they do not contain any potentially harmful genes. The SATACs are generated, not from the minichromosome fragment as, for example, in U.S. Pat. No. 5,288,625, but from the fragment of the formerly dicentric chromosome. In addition, methods for generating euchromatic minichromosomes and the use thereof are also provided herein. Methods for generating one type of MAC, the minichromosome, previously described in U.S. Pat. No. 5,288,625, and the use thereof for expression of heterologous DNA are provided. Cell lines containing the minichromosome and the use thereof for cell fusion are also provided.

In one embodiment, a cell line containing the mammalian minichromosome is used as recipient cells for donor DNA encoding a selected gene or multiple genes. To facilitate integration of the donor DNA into the minichromosome, the recipient cell line preferably contains the minichromosome but does not also contain the formerly dicentric chromosome. This may be accomplished by methods disclosed herein such as cell fusion and selection of cells that contain a minichromosome and no formerly dicentric chromosome. The donor DNA is linked to a second selectable marker and is targeted to and integrated into the minichromosome. The resulting chromosome is transferred by cell fusion into an appropriate recipient cell line, such as a Chinese hamster cell line [CHO]. After large-scale production of the cells carrying the engineered chromosome, the chromosome is isolated. In particular, metaphase chromosomes are obtained, such as by addition of colchicine, and they are purified from the cell lysate. These chromosomes are used for cloning, sequencing and for delivery of heterologous DNA into cells.

Also provided are SATACs of various sizes that are formed by repeated culturing under selective conditions and subcloning of cells that contain chromosomes produced from the formerly dicentric chromosomes. The exemplified SATACs are based on repeating NA units that are about 15 Mb [two about 7.5 Mb blocks]. The repeating DNA unit of SATACs formed from other species and other chromosomes may vary, but typically would be on the order of about 7 to about 20 Mb. The repeating DNA units are referred to herein as megareplicons, which in the exemplified SATACs contain tandem blocks of satellite DNA flanked by non-satellite DNA, including heterologous DNA and non-satellite DNA. Amplification produces an array of chromosome segments [each called an amplicon] that contain two inverted megareplicons bordered by heterologous ["foreign"] DNA. Repeated cell fusion, growth on selective medium and/or BrdU [5-bromodeoxyuridine] treatment or other treatment with other genome destabilizing reagent or agent, such as ionizing radiation, including X-rays, and subcloning results in cell lines that carry stable heterochromatic or partially heterochromatic chromosomes, including a 150-200 Mb "sausage" chromosome, a 500-1000 Mb gigachromosome, a stable 250-400 Mb megachromosome and various smaller stable chromosomes derived therefrom. These chromosomes are based on these repeating units and can include heterologous DNA that is expressed.

Thus, methods for producing MACs of both types (i.e., SATACS and minichromosomes) are provided. These methods are applicable to the production of artificial chromosomes containing centromeres derived from any higher eukaryotic cell, including mammals, irds, fowl, fish, insects and plants.

The resulting chromosomes can be purified by methods provided herein to provide vectors for introduction of heterologous DNA into selected cells for production of the gene product(s) encoded by the heterologous DNA, for production of transgenic animals, birds, fowl, fish and plants or for gene therapy.

In addition, methods and vectors for fragmenting the minichromosomes and SATACs are provided. Such methods and vectors can be used for in vivo generation of smaller stable artificial chromosomes. Vectors for chromosome fragmentation are used to produce an artificial chromosome that contains a megareplicon, a centromere and two telomeres and will be between about 7.5 Mb and about 60 Mb, preferably between about 10 Mb-15 Mb and 30-50 Mb. As exemplified herein, the preferred range is between about 7.5 Mb and 50 Mb. Such artificial chromosomes may also be produced by other methods.

Isolation of the 15 Mb [or 30 Mb amplicon containing two 15 Mb inverted repeats] or a 30 Mb or higher multimer, such as 60 Mb, thereof should provide a stable chromosomal vector that can be manipulated in vitro. Methods for reducing the size of the MACs to generate smaller stable self-replicating artificial chromosomes are also provided.

Methods and vectors for targeting heterologous DNA into the artificial chromosomes are also provided as are methods and vectors for fragmenting the chromosomes to produce smaller but stable and self- replicating artificial chromosomes.

The chromosomes are introduced into cells to produce stable transformed cell lines or cells, depending upon the source of the cells. Introduction is effected by any suitable method including, but not limited to electroporation, direct uptake, such as by calcium phosphate precipitation, uptake of isolated chromosomes by lipofection, by microcell fusion, by lipid-mediated carrier systems or other suitable method. The resulting cells can be used for production of proteins in the cells. The chromosomes can be isolated and used for gene delivery.

Methods for isolation of the chromosomes based on the DNA content of the chromosomes, which differs in MACs versus the authentic chromosomes, are provided.

These artificial chromosomes can be used in gene therapy, gene product production systems, production of humanized genetically transformed animal organs, production of transgenic plants and animals, including mammals, birds, fowl, fish, invertebrates, vertebrate, reptiles and insects, any organism or device that would employ chromosomal elements as information storage vehicles, and also for analysis and study of centromere function, for the production of artificial chromosome vectors that can be constructed in vitro, and for the preparation of species-specific artificial chromosomes. The artificial chromosomes can be introduced into cells using microinjection, cell fusion, microcell fusion, electroporation, electrofusion, projectile bombardment, calcium phosphate precipitation, site-specific targeting, lipid-mediated transfer systems and other such methods. Cells particularly suited for use with the artificial chromosomes include, but are not limited to plant cells, particularly tomato, arabidopsis, and others, insect cells, including silk worm cells, insect larvae, fish, reptiles, amphibians, arachnids, mammalian cells, avian cells, embryonic stem cells, haematopoietic stem cells, embryos and cells for use in methods of genetic therapy, such as lymphocytes that are used in methods of adoptive immunotherapy and nerve or neural cells. Thus methods of producing gene products and transgenic animals and plants are provided. Also provided are the resulting transgenic animals and plants.

Exemplary cell lines that contain these chromosomes are also provided.

Methods for preparing artificial chromosomes for particular species and for cloning centromeres are also provided. For example, two methods for generating artificial chromosomes for use in different species are provided. First, the methods herein may applied to different species. Second, means for generating species-specific artificial chromosomes and for cloning centromeres are provided. In particular, a method for cloning a centromere from an animal or plant by preparing a library of DNA fragments that contain the genome of the plant or animal, introducing each of the fragments into a mammalian satellite artificial chromosome [SATAC] that contains a centromere from a different species, generally a mammal, from the selected plant or animal, generally a non- mammal, and a selectable marker. The selected plant or animal is one in which the mammalian species centromere does not function. Each of the SATACs is introduced into the cells, which are grown under selective conditions, and cells with SATACs are identified. Such SATACS should contain a centromere encoded by the DNA from the library or should contain the necessary elements for stable replication in the selected species.

Also provided are libraries in which the relatively large fragments of DNA are contained on artificial chromosomes.

Transgenic animals, invertebrates and vertebrates, plants and insects, fish, reptiles, amphibians, arachnids, birds, fowl, and mammals are also provided. Of particular interest are transgenic animals that express genes that confer resistance or reduce susceptibility to disease. Since multiple genes can be introduced on a MAC, a series of genes encoding an antigen can be introduced, which upon expression will serve to immunize [in a manner similar to a multivalent vaccine] the host animal against the diseases for which exposure to the antigens provide immunity or some protection.

Also of interest are transgenic animals that serve as models of certain diseases and disorders for use in studying the disease and developing therapeutic treatments and cures thereof. Such animal models of disease express genes [typically carrying a disease-associated mutation], which are introduced into the animal on a MAC and which induce the disease or disorder in the animal. Similarly, MACs carrying genes encoding antisense RNA may be introduced into animal cells to generate conditional "knock-out" transgenic animals. In such animals, expression of the antisense RNA results in decreased or complete elimination of the products of genes corresponding to the antisense RNA. Of further interest are transgenic mammals that harbor MAC-carried genes encoding therapeutic proteins that are expressed in the animal's milk. Transgenic animals for use in xenotransplantation, which express MAC- carried genes that serve to humanize the animal's organs, are also of interest. Genes that might be used in humanizing animal organs include those encoding human surface antigens.

Methods for cloning centromeres, such as mammalian centromeres, are also provided. In particular, in one embodiment, a library composed of fragments of SATACs are cloned into YACs [yeast artificial chromosomes] that include a detectable marker, such as DNA encoding tyrosinase, and then introduced into mammalian cells, such as albino mouse embryos. Mice produced from embryos containing such YACs that include a centromere that functions in mammals will express the detectable marker. Thus, if mice are produced from albino mouse embryos into which a functional mammalian centromere was introduced, the mice will be pigmented or have regions of pigmentation.

3.6.4 Particularly Relevant Definitions

Unless defined otherwise, all technical and scientific terms used herein have the same meaning as is commonly understood by one of skill in the art to which this invention belongs. All patents and publications referred to herein are incorporated by reference.

As used herein, a mammalian artificial chromosome [MAC] is a piece of DNA that can stably replicate and segregate alongside endogenous chromosomes. It has the capacity to accommodate and express heterologous genes inserted therein. It is referred to as a mammalian artificial chromosome because it includes an active mammalian centromere(s). Plant artificial chromosomes, insect artificial chromosomes and avian artificial chromosomes refer to chromosomes that include plant and insect centromeres, respectively. A human artificial chromosome [HAC] refers to chromosomes that include human centromeres, BUGACs refer to insect artificial chromosomes, and AVACs refer to avian artificial chromosomes.

As used herein, stable maintenance of chromosomes occurs when at least about 85%, preferably 90%, more preferably 95%, of the cells retain the chromosome. Stability is measured in the presence of a selective agent. Preferably these chromosomes are also maintained in the absence of a selective agent. Stable chromosomes also retain their structure during cell culturing, suffering neither intrachromosomal nor interchromosomal rearrangements.

As used herein, growth under selective conditions means growth of a cell under conditions that require expression of a selectable marker for survival.

As used herein, euchromatin and heterochromatin have their recognized meanings, euchromatin refers to DNA that contains genes, and heterochromatin refers to chromatin that has been thought to be inactive. Highly repetitive DNA sequences [satellite DNA], at least with respect to mammalian cells, are usually located in regions of centromeric heterochromatin [pericentric heterochromatin]. Constitutive heterochromatin refers to heterochromatin that contains the highly repetitive DNA which is constitutively condensed and genetically inactive.

As used herein, BrdU refers to 5-bromodeoxyuridine, which during replication is inserted in place of thymidine. BrdU is used as a mutagen; it also inhibits condensation of metaphase chromosomes during cell division.

As used herein, a dicentric chromosome is a chromosome that contains two centromeres. A multicentric chromosome contains more than two centromeres.

As used herein, a formerly dicentric chromosome is a chromosome that is produced when a dicentric chromosome fragments and acquires new telomeres so that two chromosomes, each having one of the centromeres, are produced. Each of the fragments are replicable chromosomes. If one of the chromosomes undergoes amplification of euchromatic DNA to produce a full functionally chromosome that contains the newly introduced heterologous DNA and primarily [at least more than 50%] euchromatin, it is a minichromosome. The remaining chromosome is a formerly dicentric chromosome. If one of the chromosomes undergoes amplification, whereby heterochromatin [satellite DNA] is amplified and a euchromatic portion [or arm] remains, it is referred to as a sausage hromosome. A chromosome that is substantially all heterochromatin, except for portions of heterologous DNA, is called a SATAC. Such chromosomes [SATACs] can be produced from sausage chromosomes by culturing the cell containing the sausage chromosome under conditions, such as BrdU treatment and/or growth under selective conditions, that destabilize the chromosome so that a satellite artificial chromosomes [SATAC] is produced. For purposes herein, it is understood that SATACs may not necessarily be produced in multiple steps, but may appear after the initial introduction of the heterologous DNA and growth under selective conditions, or they may appear after several cycles of growth under selective conditions and BrdU treatment.

As used herein an amplicon is a repeated DNA amplification unit that contains a set of inverted repeats of the megareplicon. A megareplicon represents a higher order replication unit. For example, with reference to the SATACs, the megareplicon contains a set of tandem DNA blocks each containing satellite DNA flanked by non-satellite DNA. Contained within the megareplicon is a primary replication site, referred to as the megareplicator, which may be involved in organizing and facilitating replication of the pericentric heterochromatin and possibly the centromeres. Within the megareplicon there may be smaller [e.g., 50- 300 kb in some mammalian cells] secondary replicons. In the exemplified SATACS, the megareplicon is defined by two tandem about 7.5 Mb DNA blocks [see, e.g., FIG. 3]. Within each artificial chromosome [AC] or among a population thereof, each amplicon has the same gross structure but may contain sequence variations. Such variations will arise as a result of movement of mobile genetic elements, deletions or insertions or mutations that arise, particularly in culture. Such variation does not affect the use of the ACs or their overall structure as described herein.

As used herein, the minichromosome refers to a chromosome derived from a multicentric, typically dicentric, chromosome [see, e.g., FIG. 1] that contains more euchromatic than heterochromatic DNA.

As used herein, a megachromosome refers to a chromosome that, except for introduced heterologous DNA, is substantially composed of heterochromatin.

Megachromosomes are made of an array of repeated amplicons that contain two inverted megareplicons bordered by introduced heterologous DNA [see, e.g., FIG. 3 for a schematic drawing of a megachromosome]. For purposes herein, a megachromosome is about 50 to 400 Mb, generally about 250-400 Mb. Shorter variants are also referred to as truncated megachromosomes [about 90 to 120 or 150 Mb], dwarf megachromosomes [about 150-200 Mb] and cell lines, and a micro-megachromosome [about 60-90 Mb]. For purposes herein, the term megachromosome refers to the overall repeated structure based on an array of repeated chromosomal segments [amplicons] that contain two inverted megareplicons bordered by any inserted heterologous DNA. The size will be specified.

As used herein, genetic therapy involves the transfer or insertion of heterologous DNA into certain cells, target cells, to produce specific gene products that are involved in correcting or modulating disease. The DNA is introduced into the selected target cells in a manner such that the heterologous DNA is expressed and a product encoded thereby is produced. Alternatively, the heterologous DNA may in some manner mediate expression of DNA that encodes the therapeutic product. It may encode a product, uch as a peptide or RNA, that in some manner mediates, directly or indirectly, expression of a therapeutic product. Genetic therapy may also be used to introduce therapeutic compounds, such as TNF, that are not normally produced in the host or that are not produced in therapeutically effective amounts or at a therapeutically useful time. Expression of the heterologous DNA by the target cells within an organism afflicted with the disease thereby enables modulation of the disease. The heterologous DNA encoding the therapeutic product may be modified prior to introduction into the cells of the afflicted host in order to enhance or otherwise alter the product or expression thereof.

As used herein, heterologous or foreign DNA and RNA are used interchangeably and refer to DNA or RNA that does not occur naturally as part of the genome in which it is present or which is found in a location or locations in the genome that differ from hat in which it occurs in nature. It is DNA or RNA that is not endogenous to the cell and has been exogenously introduced into the cell. Examples of heterologous DNA include, but are not limited to, DNA that encodes a gene product or gene product(s) of interest, introduced for purposes of gene therapy or for production of an encoded protein. Other examples of heterologous DNA include, but are not limited to, DNA that encodes traceable marker proteins, such as a protein that confers drug resistance, DNA that encodes therapeutically effective substances, such as anti-cancer agents, enzymes and hormones, and DNA that encodes other types of proteins, such as antibodies. Antibodies that are encoded by heterologous DNA may be secreted or expressed on the surface of the cell in which the heterologous DNA has been introduced.

As used herein, a therapeutically effective product is a product that is encoded by heterologous DNA that, upon introduction of the DNA into a host, a product is expressed that effectively ameliorates or eliminates the symptoms, manifestations of an inherited or acquired disease or that cures said disease.

As used herein, transgenic plants refer to plants in which heterologous or foreign DNA is expressed or in which the expression of a gene naturally present in the plant has been altered.

As used herein, operative linkage of heterologous DNA to regulatory and effector sequences of nucleotides, such as promoters, enhancers, transcriptional and translational stop sites, and other signal sequences refers to the relationship between such DNA and such sequences of nucleotides. For example, operative linkage of heterologous DNA to a promoter refers to the physical relationship between the DNA and the promoter such that the transcription of such DNA is initiated from the promoter by an RNA polymerase that specifically recognizes, binds to and transcribes the DNA in reading frame.

As used herein, isolated, substantially pure DNA refers to DNA fragments purified according to standard techniques employed by those skilled in the art, such as that found in Maniatis et al. [(1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.].

As used herein, expression refers to the process by which nucleic acid is transcribed into mRNA and translated into peptides, polypeptides, or proteins. If the nucleic acid is derived from genomic DNA, expression may, if an appropriate eukaryotic host ell or organism is selected, include splicing of the mRNA.

As used herein, vector or plasmid refers to discrete elements that are used to introduce heterologous DNA into cells for either expression of the heterologous DNA or for replication of the cloned heterologous DNA. Selection and use of such vectors and lasmids are well within the level of skill of the art.

As used herein, transformation/transfection refers to the process by which DNA or RNA is introduced into cells. Transfection refers to the taking up of exogenous nucleic acid, e.g., an expression vector, by a host cell whether or not any coding sequences are in fact expressed. Numerous methods of transfection are known to the ordinarily skilled artisan, for example, by direct uptake using calcium phosphate [CaPO4; see, e.g., Wigler et al. (1979) Proc. Natl. Acad. Sci. U.S.A. 76:1373- 1376], polyethylene glycol [PEG]-mediated DNA uptake, electroporation, lipofection [see, e.g., Strauss (1996) Meth. Mol. Biol. 54:307-327], microcell fusion [see, EXAMPLES, see, also Lambert (1991) Proc. Natl. Acad. Sci. U.S.A. 88:5907-5911; U.S. Pat. No. 5,396,767, Sawford et al. (1987) Somatic Cell Mol. Genet. 13:279-284; Dhar et al. (1984) Somatic Cell Mol. Genet. 10:547-559; and McNeill-Killary et al. (1995) Meth. Enzymol. 254:133-152], lipid-mediated carrier systems [see, e.g., Teifel et al. (1995) Biotechnigues 19:79-80; Albrecht et al. (1996) Ann. Hematol. 72:73-79; Holmen et al. (1995) In Vitro Cell Dev. Biol. Anim. 31:347-351; REmy et al. (1994) Bioconjug. Chem. 5:647-654; Le Bolch et al. (1995) Tetrahedron Lett. 36:6681-6684; Loeffler et al. (1993) Meth. Enzymol. 217:599-618] or other suitable method. Successful transfection is generally recognized by detection of the presence of the heterologous nucleic acid within the transfected cell, such as any indication of the operation of a vector within the host cell. Transformation means introducing DNA into an organism so that the DNA is replicable, either as an extrachromosomal element or by chromosomal integration.

As used herein, injected refers to the microinjection [use of a small syringe] of DNA into a cell.

As used herein, substantially homologous DNA refers to DNA that includes a sequence of nucleotides that is sufficiently similar to another such sequence to form stable hybrids under specified conditions.

It is well known to those of skill in this art that nucleic acid fragments with different sequences may, under the same conditions, hybridize detectably to the same "target" nucleic acid. Two nucleic acid fragments hybridize detectably, under stringent conditions over a sufficiently long hybridization period, because one fragment contains a segment of at least about 14 nucleotides in a sequence which is complementary [or nearly complementary] to the sequence of at least one segment in the other nucleic acid fragment. If the time during which hybridization is allowed to occur is held constant, at a value during which, under preselected stringency conditions, two nucleic acid fragments with exactly complementary base-pairing segments hybridize detectably to each other, departures from exact complementarity can be introduced into the base-pairing segments, and base-pairing will nonetheless occur to an extent sufficient to make hybridization detectable. As the departure from complementarity between the base- pairing segments of two nucleic acids becomes larger, and as conditions of the hybridization become more stringent, the probability decreases that the two segments will hybridize detectably to each other.

Two single-stranded nucleic acid segments have "substantially the same sequence," within the meaning of the present specification, if (a) both form a base-paired duplex with the same segment, and (b) the melting temperatures of said two duplexes in a solution of 0.5.times. SSPE differ by less than 10.degree. C. If the segments being compared have the same number of bases, then to have "substantially the same sequence", they will typically differ in their sequences at fewer than 1 base in 10. Methods for determining melting temperatures of nucleic acid duplexes are well known [see, e.g., Meinkoth and Wahl (1984) Anal. Biochem. 138:267-284 and references cited therein].

As used herein, a nucleic acid probe is a DNA or RNA fragment that includes a sufficient number of nucleotides to specifically hybridize to DNA or RNA that includes identical or closely related sequences of nucleotides. A probe may contain any number of nucleotides, from as few as about 10 and as many as hundreds of thousands of nucleotides. The conditions and protocols for such hybridization reactions are well known to those of skill in the art as are the effects of probe size, temperature, degree of ismatch, salt concentration and other parameters on the hybridization reaction. For example, the lower the temperature and higher the salt concentration at which the hybridization reaction is carried out, the greater the degree of mismatch that may be present in the hybrid molecules.

To be used as a hybridization probe, the nucleic acid is generally rendered detectable by labelling it with a detectable moiety or label, such as .sup.32 p, .sup.3 H and .sup.14 C, or by other means, including chemical labelling, such as by nick-translation in the presence of deoxyuridylate biotinylated at the 5'-position of the uracil moiety. The resulting probe includes the biotinylated uridylate in place of thymidylate residues and can be detected [via the biotin moieties] by any of a number of commercially available detection systems based on binding of streptavidin to the biotin. Such commercially available detection systems can be obtained, for example, from Enzo Biochemicals, Inc. [New York, N.Y.]. Any other label known to those of kill in the art, including non-radioactive labels, may be used as long as it renders the probes sufficiently detectable, which is a function of the sensitivity of the assay, the time available [for culturing cells, extracting DNA, and hybridization assays], the quantity of DNA or RNA available as a source of the probe, the particular label and the means used to detect the label.

Once sequences with a sufficiently high degree of homology to the probe are identified, they can readily be isolated by standard techniques, which are described, for example, by Maniatis et al. ((1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.).

As used herein, conditions under which DNA molecules form stable hybrids and are considered substantially homologous are such that DNA molecules with at least about 60% complementarity form stable hybrids. Such DNA fragments are herein considered to be "substantially homologous". For example, DNA that encodes a particular protein is substantially homologous to another DNA fragment if the DNA forms stable hybrids such that the sequences of the fragments are at least about 60% complementary and if a protein encoded by the DNA retains its activity.

For purposes herein, the following stringency conditions are defined:

1) high stringency: 0.1.times.SSPE, 0.1% SDS, 65.degree. C.

2) medium stringency: 0.2.times.SSPE, 0.1% SDS, 50.degree. C.

3) low stringency: 1.0 x SSPE, 0.1% SDS, 50.degree. C.

or any combination of salt and temperature and other reagents that result in selection of the same degree of mismatch or matching.

As used herein, immunoprotective refers to the ability of a vaccine or exposure to an antigen or immunity-inducing agent, to confer upon a host to whom the vaccine or antigen is administered or introduced, the ability to resist infection by a disease-causing pathogen or to have reduced symptoms. The selected antigen is typically an antigen that is presented by the pathogen.

As used herein, all assays and procedures, such as hybridization reactions and antibody-antigen reactions, unless otherwise specified, are conducted under conditions recognized by those of skill in the art as standard conditions.

3.6.5 Preparation of cell lines containing MACs

The methods, cells and MACs provided herein are produced by virtue of the discovery of the existence of a higher-order replication unit [megareplicon] of the centromeric region. This megareplicon is delimited by a primary replication initiation site [megareplicator], and appears to facilitate replication of the centromeric heterochromatin, and most likely, centromeres. Integration of heterologous DNA into the megareplicator region or in close proximity thereto, initiates a large-scale amplification of megabase-size chromosomal segments, which leads to de novo chromosome formation in living cells.

Cell lines containing MACs can be prepared by transforming cells, preferably a stable cell line, with a heterologous DNA fragment that encodes a selectable marker, culturing under selective conditions, and identifying cells that have a multicentric, typically dicentric, chromosome. These cells can then be manipulated as described herein to produce the minichromosomes and other MACs, particularly the heterochromatic SATACs as described herein.

Development of a multicentric, particularly dicentric, chromosome typically is effected through integration of the heterologous DNA in the pericentric heterochromatin. Thus, the probability of incorporation can be increased by including DNA, such as satellite DNA, in the heterologous fragment that encodes the selectable marker. The resulting cell lines can then be treated as the exemplified cells herein to produce cells in which the dicentric chromosome has fragmented and to introduce additional selective markers into the dicentric chromosome, whereby amplification of the pericentric heterochromatin will produce the heterochromatic chromosomes. The following discussion is with reference to the EC3/7 line and use of resulting cells. The same procedures can be applied to any other cells, particularly cell lines to create SATACs and euchromatic minichromosomes.

3.6.5.1 Formation of de novo chromosomes

De novo centromere formation in a transformed mouse LMTK-fibro- blast cell line [EC3/7] after cointegration of lambda constructs [lambda CM8 and lambda gtWESneo] carrying human and bacterial DNA [Hadlaczky et al. (1991) Proc. Natl. Acad. Sci. U.S.A. 88:8106-8110 and U.S. application Ser. No. 08/375,271] has been shown. The integration of the "heterologous" engineered human, bacterial and phage DNA, and the subsequent amplification of mouse and heterologous DNA that led to the formation of a dicentric chromosome, occurred at the centromeric region of the short arm of a mouse chromosome. By G-banding, this chromosome was identified as mouse chromosome 7. Because of the presence of two functionally active centromeres on the same chromosome, regular breakages occur between the centromeres. Such specific chromosome breakages gave rise to the appearance [in approximately 10% of the cells] of a chromosome fragment carrying the neo-centromere. From the EC3/7 cell line [see, U.S. Pat. No. 5,288,625, deposited at the European Collection of Animal Cell Culture (hereinafter ECACC) under accession no. 90051001; see, also Hadlaczky et al. (1991) Proc. Natl. Acad. Sci. U.S.A. 88:8106-8110, and U.S. application Ser. No. 08/375,271 and the corresponding published European application EP 0 473 253, two sublines [EC3/7C5 and EC3/7C6] were selected by repeated single-cell cloning. In these cell lines, the neo-centromere was found exclusively on a minichromosome [neo-minichromosome], while the formerly dicentric chromosome carried traces of "heterologous" DNA.

It has now been discovered that integration of DNA encoding a selectable marker in the heterochromatic region of the centromere led to formation of the dicentric chromosome.

3.6.5.2 The neo-minichromosome

The chromosome breakage in the EC3/7 cells, which separates the neo-centromere from the mouse chromosome, occurred in the G-band positive "heterologous" DNA region. This is supported by the observation of traces of lambda and human DNA sequences at the broken end of the formerly dicentric chromosome. Comparing the G-band pattern of the chromosome fragment carrying the neo-centromere with that of the stable neo-minichromosome, it is apparent that the neo-minichromosome is an inverted duplicate of the chromosome fragment that bears the neo-centromere. This is supported by the observation that although the neo-minichromosome carries only one functional centromere, both ends of the minichromosome are heterochromatic, and mouse satellite DNA sequences were found in these heterochromatic regions by in situ hybridization.

Mouse cells containing the minichromosome, which contains multiple repeats of the heterologous DNA, which in the exemplified embodiment is lambda DNA and the neomycin-resistance gene, can be used as recipient cells in cell transformation. Donor DNA, such as selected heterologous DNA containing lambda DNA linked to a second selectable marker, such as the gene encoding hygromycin phosphotransferase which confers hygromycin resistance [hyg], can be introduced into the mouse cells and integrated into the minichromosomes by homologous recombination of lambda

DNA in the donor DNA with that in the minichromosomes. Integration is verified by in situ hybridization and Southern blot analyses. Transcription and translation of the heterologous DNA is confirmed by primer extension and immunoblot analyses.

For example, DNA has been targeted into the lambda neo-minichromosome in EC3/7C5 cells using a lambda DNA-containing construct [pNem1ruc] that also contains DNA encoding hygromycin resistance and the Renilla luciferase gene linked to a promoter, such as the cytomegalovirus [CMV] early promoter, and the bacterial neomycin resistance-encoding DNA. Integration of the donor DNA into the chromosome in selected cells [designated PHN4] was confirmed by nucleic acid amplification [PCR] and in situ hybridization.

The resulting engineered minichromosome that contains the heterologous DNA can then be transferred by cell fusion into a recipient cell line, such as Chinese hamster ovary cells [CHO] and correct expression of the heterologous DNA can be verified. Following production of the cells, metaphase chromosomes are obtained, such as by addition of colchicine, and the chromosomes purified by addition of AT and GC specific dyes on a dual laser beam based cell sorter. Preparative amounts of chromosomes [5×10^4-5×10^7 chromosomes/ml] at a purity of 95% or higher can be obtained. The resulting chromosomes are used for delivery to cells by methods such as microinjection and liposome-mediated transfer.

Thus, the neo-minichromosome is stably maintained in cells, replicates autonomously, and permits the persistent long-term expression of the neo gene under non-selective culture conditions. It also contains megabases of heterologous known DNA [lambda NA in the exemplified embodiments] that serves as target sites for homologous recombination and integration of DNA of interest. The neo-minichromosome is, thus, a vector for genetic engineering of cells.

The methods herein provide means to induce the events that lead to formation of the neo-minichromosome by introducing heterologous DNA with a selective marker [preferably a dominant selectable marker] into cells and culturing the cells under selective conditions. As a result, cells that contain a multicentric, e.g., dicentric chromosome, or fragments thereof, generated by amplification are produced. Cells with the dicentric chromosome can then be treated to destabilize the chromosomes with agents, such as BrdU and/or culturing under selective conditions, resulting in cells in which the dicentric chromosome has formed two chromosomes, a so-called minichromosome, and a formerly dicentric chromosome that has typically undergone amplification in the heterochromatin where the heterologous DNA has integrated to produce a SATAC or a sausage chromosome [discussed below]. These cells can be fused with other cells to separate the minichromosome from the formerly dicentric chromosome into different cells so hat each type of MAC can be manipulated separately.

3.6.5.3 Preparation of SATACs

To prepare a SATAC, the starting materials are cells, preferably a stable cell line, such as a fibroblast cell line, and a DNA fragment that includes DNA that encodes a selective marker. The DNA fragment is introduced into the cell by methods of DNA transfer, including but not limited to direct uptake using calcium phosphate, electroporation, and lipid-mediated transfer. To insure integration of the DNA fragment in the heterochromatin, it is preferable to start with DNA that will be targeted to the pericentric heterochromatic region of the chromosome, such as lambda CM8 and vectors provided herein, such as pTEMPUD that include satellite DNA. After introduction of the DNA, the cells are grown under selective conditions. The resulting cells are examined and any that have multicentric, particularly dicentric, chromosomes, or heterochromatic chromosomes or sausage chromosomes or other such structure are selected.

In particular, if a cell with a dicentric chromosome is selected, it can be grown under selective conditions, or, preferably, additional DNA encoding a second selectable marker is introduced, and the cells grown under conditions selective for the second marker. Cells with a structure, such as the sausage chromosome, can be selected and fused with a second cell line to eliminate other chromosomes that are not of interest. If desired, cells with other chromosomes can be selected and treated as described herein. If a cell with a sausage chromosome is selected, it can be treated with an agent, such as BrdU, that destabilizes the chromosome so that the heterochromatic arm forms a chromosome that is substantially heterochromatic [i.e., a megachromosome]. Structures such as the gigachromsome in which the heterochromatic arm has amplified but not broken off from the euchromatic arm, will also be observed. The megachromosome is a stable chromosome. Further manipulation, such as fusions and growth in selective conditions and/or BrdU treatment or other such treatment, can lead to fragmentation of the egachromosome to form smaller chromosomes that have the amplicon as the basic repeating unit.

The megachromosome can be further fragmented in vivo using a chromosome fragmentation vector, such as pTEMPUD to ultimately produce a chromosome that comprises a smaller stable replicable unit, about 15 Mb-60 Mb, containing one to four megareplicons.

Thus, the stable chromosomes formed de novo that originate from the short arm of mouse chromosome 7 have been analyzed. This chromosome region shows a capacity for amplification of large chromosome segments, and promotes de novo chromosome formation. Large-scale amplification at the same chromosome region leads to the formation of dicentric and multicentric chromosomes, a minichromosome, the 150-200 Mb size lambda neo-chromosome, the "sausage" chromosome, the 500-1000 Mb gigachromosome, and the stable 250-400 Mb megachromosome.

A clear segmentation is observed along the arms of the megachromosome, and analyses show that the building units of this chromosome are amplicons of ~30 Mb composed of mouse major satellite DNA with the integrated "foreign" DNA sequences at both ends. The ~30 Mb amplicons are composed of two ~15 Mb inverted doublets of ~7.5 Mb mouse major satellite DNA blocks, which are separated from each other by a narrow band of non-satellite sequences. The wider non-satellite regions at the amplicon borders contain integrated, exogenous [heterologous] DNA, while the narrow bands of non-satellite DNA sequences within the amplicons are integral parts of the pericentric heterochromatin of mouse chromosomes. These results indicate that the ~7.5 Mb blocks flanked by non- satellite DNA are the building units of the pericentric heterochromatin of mouse chromosomes, and the ~15 Mb size pericentric regions of mouse chromosomes contain two ~7.5 Mb units.

Apart from the euchromatic terminal segments, the whole megachromosome is heterochromatic, and has structural homogeneity. Therefore, this large chromosome offers a unique possibility for obtaining information about the amplification process, and for analyzing some basic characteristics of the pericentric constitutive heterochromatin, as a vector for heterologous DNA, and as a target for further fragmentation.

As shown herein, this phenomenon is generalizable and can be observed with other chromosomes. Also, although these de novo formed chromosome segments and chromosomes appear different, there are similarities that indicate that a similar amplification mechanism plays a role in their formation: (i) in each case, the amplification is initiated in the centromeric region of the mouse chromosomes and large (Mb size) amplicons are formed; (ii) mouse major satellite DNA sequences are constant constituents of the amplicons, either by providing the bulk of the heterochromatic amplicons [H-type amplification], or by bordering the euchromatic amplicons [E-type amplification]; (iii) formation of inverted segments can be demonstrated in the lambda neo-chromosome and megachromosome; (iv) chromosome arms and chromosomes formed by the amplification are stable and functional.

The presence of inverted chromosome segments seems to be a common phenomenon in the chromosomes formed de novo at the centromeric region of mouse chromosome 7. During the formation of the neo-minichromosome, the event leading to the stabilization of the distal segment of mouse chromosome 7 that bears the neo-centromere may have been the formation of its inverted duplicate. Amplicons of the megachromosome are inverted doublets of ~7.5 Mb mouse major satellite DNA blocks.

3.6.5.4 Cell lines

Cell lines that contain MACs, such as the minichromosome, the . lambda.-neo chromosome, and the SATACs are provided herein or can be produced by the methods herein. Such cell lines provide a convenient source of these chromosomes and can be manipulated, such as by cell fusion or production of microcells for fusion with selected cell lines, to deliver the chromosome of interest into hybrid cell lines. Exemplary cell lines are described herein and some have been deposited with the ECACC.

3.6.5.4.1 EC3/7C5 and EC3/7C6

Cell lines EC3/7C5 and EC3/7C6 were produced by single cell cloning of EC3/7. For exemplary purposes EC3/7C5 has been deposited with the ECACC. These cell lines contain a minichromosome and the formerly dicentric chromosome from EC3/7. The stable mini-chromosomes in cell lines EC3/7C5 and EC3/7C6 appear to be the same and they seem to be duplicated derivatives of the 10-15 Mb "broken-off" fragment of the dicentric chromosome. Their similar size in these independently generated cell lines might indicate that about 20-30 Mb is the minimal or close to the minimal physical size for a stable minichromosome.

3.6.5.4.2 TF1004G19

Introduction of additional heterologous DNA, including DNA encoding a second selectable marker, hygromycin phosphotransferase, i.e., the hygromycin-resistance gene, and also a detectable marker, beta-galactosidase (i.e., encoded by the lacz gene), into the EC3/7C5 cell line and growth under selective conditions produced cells designated TF1004G19. In particular, this cell line was produced from the EC3/7C5 cell line by cotransfection with plasmids pH132, which contains an anti-HIV ribozyme and hygromycin-resistance gene, pCH110 [encodes, beta-galactosidase] and lambda phage [lambdacl 875 Sam 7] DNA and selection with hygromycin B.

Detailed analysis of the TF1004G19 cell line by in situ hybridization with lambda phage and plasmid DNA sequences revealed the formation of the sausage chromosome. The formerly dicentric chromosome of the EC3/7C5 cell line translocated to the end of another acrocentric chromosome. The heterologous DNA integrated into the pericentric heterochromatin of the formerly dicentric chromosome and is amplified several times with megabases of mouse pericentric heterochromatic satellite DNA sequences forming the "sausage" chromosome. Subsequently the acrocentric mouse chromosome was substituted by a euchromatic telomere.

In situ hybridization with biotin-labeled subfragments of the hygromycin-resistance and, beta-galactosidase genes resulted in a hybridization signal only in the heterochromatic arm of the sausage chromosome, indicating that in TF1004G19 transformant cells these genes are localized in the pericentric heterochromatin. A high level of gene expression, however, was detected.

In general, heterochromatin has a silencing effect in Drosophila, yeast and on the HSV-tk gene introduced into satellite DNA at the mouse centromere. Thus, it was of interest to study the TF1004G19 transformed cell line to confirm that genes located in the heterochromatin were indeed expressed, contrary to recognized dogma.

For this purpose, subclones of TF1004G19, containing a different sausage chromosome, were established by single cell cloning. Southern hybridization of DNA isolated from the subclones with subfragments of hygromycin phosphotransferase and lacZ genes showed a close correlation between the intensity of hybridization and the length of the sausage chromosome. This finding supports the conclusion that these genes are localized in the heterochromatic arm of the sausage chromosome.

3.6.5.4.2.1 TF1004G-19C5

TF1004G-19C5 is a mouse LMTK- fibroblast cell line containing neo-minichromosomes and stable "sausage" chromosomes. It is a subclone of TF1004G19 and was generated by single-cell cloning of the TF1004G19 cell line. It has been deposited with the ECACC as an exemplary cell line and exemplary source of a sausage chromosome. Subsequent fusion of this cell line with CHO K20 cells and selection with hygromycin and G418 and HAT (hypoxanthine, aminopteria, and thymidine medium, see Szybalski et al. (1962) Proc. Natl. Acad. Sci. 48:2026) resulted in hybrid cells (designated 19C5xHa4) that carry the sausage chromosome and/or the neo- minichromosome. BrdU treatment of the hybrid cells, followed by single cell cloning and selection with G418 and/or hygromycin produced various cells that carry chromosomes of interest, including G43 and G3D5.

3.6.5.4.2.2 Other Subclones

Cell lines GB43 and G3D5 were obtained by treating 19C5xHa4 cells with BrdU followed by growth in G418-containing selective medium and retreatment with BrdU. The two cell lines were isolated by single cell cloning of the selected cells. GB43 cells contain the neo-minichromosome only. G3D5, which has been deposited with the ECACC, carries the neo- minichromosome and the megachromosome. Single cell cloning of this cell line followed by growth of the subclones in G418- and hygromycin- containing medium yielded subclones such as the GHB42 cell line carrying the neo-minichromosome and the megachromosome. H1D3 is a mouse-hamster hybrid cell line carrying the megachromosome, but no neo-minichromosome, and was generated by treating 19C5xHa4 cells with rdU followed by growth in hygromycin-containing selective medium and single cell subcloning of selected cells. Fusion of this cell line with the CD4$^+$HeLa cell line that also carries DNA encoding an additional selection gene, the neomycin-resistance gene, produced cells [designated H1xHE41 cells] that carry the megachromosome as well as a human chromosome that carries CD4neo [H1D3 cells]. Further BrdU treatment and single cell cloning produced cell lines, such as 1B3, that include cells with a truncated megachromosome.

3.6.5.5 DNA constructs used to transform the cells

Heterologous DNA can be introduced into the cells by transfection or other suitable method at any stage during preparation of the chromosomes. In general, incorporation of such DNA into the MACs is assured through site-directed integration, such as may be accomplished by inclusion of lambda-DNA in the heterologous DNA (for the exemplified chromosomes), and also an additional selective marker gene. For example, cells containing a MAC, such as the minichromosome or a SATAC, can be cotransfected with a plasmid carrying the desired heterologous DNA, such as DNA encoding an HIV ribozyme, the cystic fibrosis gene, and DNA encoding a second selectable marker, such as hygromycin resistance. Selective pressure is then applied to the cells by exposing them to an agent that is harmful to cells that do not express the new selectable marker. In this manner, cells that include the heterologous DNA in the MAC are identified. Fusion with a second cell line can provide a means to produce cell lines that contain one particular type of chromosomal structure or MAC.

Various vectors for this purpose can be readily constructed. The vectors preferably include DNA that is homologous to DNA contained within a MAC in order to target the DNA to the MAC for integration therein. The vectors also include a selectable marker gene and the selected heterologous gene(s) of interest. Based on the disclosure herein and the knowledge of the skilled artisan, one of skill can construct such vectors.

Of particular interest herein is the vector pTEMPUD and derivatives thereof that can target DNA into the heterochromatic region of selected chromosomes. These vectors can also serve as fragmentation vectors.

Heterologous genes of interest include any gene that encodes a therapeutic product and DNA encoding gene products of interest. These genes and DNA include, but are not limited to: the cystic fibrosis gene [CF], the cystic fibrosis transmembrane regulator (CFTR) gene [see, e.g., U.S. Pat. No. 5,240,846; Rosenfeld et al. (1992) Cell 68:143-155; Hyde et al. (1993) Nature 362: 250-255; Kerem et al. (1989) Science 245:1073- 1080; Riordan et al.(1989) Science 245:1066-1072; Rommens et al. (1989) Science 245:1059-1065; Osborne et al. (1991) Am. J. Hum. Genetics 48:6089-6122; White et al. (1990) Nature 344:665-667; Dean et al. (1990) Cell 61:863-870; Erlich et al. (1991) Science 252:1643; and U.S. Pat. Nos. 5,453,357, 5,449,604, 5,434,086, and 5,240,846, which provides a retroviral vector encoding the normal CFTR gene].

3.6.6 Isolation of artificial chromosomes

The MACs provided herein can be isolated by any suitable method known to those of skill in the art. Also, a method is provided herein for effecting substantial purification, particularly of the SATACs. SATACs have been isolated by fluorescence-activated cell sorting [FACS]. This method takes advantage of the nucleotide base content of the SATACs, which, by virtue of their heterochromatic DNA content, will differ from any other chromosomes in a cell. In particular, metaphase chromosomes are isolated (e.g., by addition of colchicine) and stained with base- specific dyes, such as Hoechst 33258 and chromomycin A3. Fluorescence- activated cell sorting will separate the SATACs from the genomic chromosomes. A dual-laser cell sorter FACStar Plus and FAXStar Vantage Becton Dickinson Immunocytometry System] in which two lasers were set to excite the dyes separately, allowed a bivariate analysis of the chromosomes by base-pair composition and size. Cells containing such SATACs can be similarly sorted.

3.6.7 Introduction of artificial chromosomes into cells, tissues, animals and plants

Suitable hosts for introduction of the MACs provided herein, include, but are not limited to, any animal or plant, cell or tissue thereof, including, but not limited to: mammals, birds, reptiles, amphibians, insects, fish, arachnids, tobacco, tomato, wheat, plants and algae. The MACs, if contained in cells, may be introduced by cell fusion or microcell fusion or, if the MACs have been isolated from cells, they may be introduced into host cells by any method known to those of skill in this art, including but not limited to: direct DNA transfer, electroporation, lipid-mediated transfer, e.g., lipofection and liposomes, microprojectile bombardment, microinjection in cells and embryos, protoplast regeneration for plants, and any other suitable method [see, e.g., Weissbach et al. (1988) Methods for Plant Molecular Biology, Academic Press, N.Y., Section VIII, pp. 421-463; Grierson et al. (1988) Plant Molecular Biology, 2d Ed., Blackie, London, Ch. 7-9; see, also U.S. Pat. Nos. 5,491,075; 5,482,928; and 5,424,409; see, also, e.g., U.S. Pat. No. 5,470,708, which describes particle-mediated transformation of mammalian unattached cells].

Other methods for introducing DNA into cells include nuclear microinjection, electroporation, and bacterial protoplast fusion with intact cells. Polycations, such as polybrene and polyornithine, may also be used. For various techniques for transforming mammalian cells, see e.g., Keown et al. Methods in Enzymology (1990) Vol. 185, pp. 527-537; and Mansour et al. (1988) Nature 336:348-352.

DNA may be introduced by direct DNA transformation; microinjection in cells or embryos, protoplast regeneration for plants, electroporation, microprojectile gun and other such methods [see, e.g., Weissbach et al. (1988) Methods for Plant Molecular Biology, Academic Press, N.Y., Section VIII, pp. 421-463; Grierson et al. (1988) Plant Molecular Biology, 2d Ed., Blackie, London, Ch. 7-9; see, also U.S. Pat. Nos. 5,491,075; 5,482,928; and 5,424,409; see, also, e.g., U.S. Pat. No. 5,470,708, which describes particle-mediated transformation of mammalian unattached cells].

For example, isolated, purified artificial chromosomes can be injected into an embryonic cell line such as a human kidney primary embryonic cell line [ATCC accession number CRL 1573] or embryonic stem cells [see, e.g., Hogan et al. (1994) Manipulating he Mouse Embryo, A :Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., see, especially, pages 255-264 and Appendix 3]. Preferably the chromosomes are introduced by microinjection, using a system such as the Eppendorf automated microinjection system, and grown under selective conditions, such as in the presence of hygromycin B or neomycin.

3.6.7.1. Methods for introduction of chromosomes into hosts

Depending on the host cell used, transformation is done using standard techniques appropriate to such cells. These methods include any, including those described herein, known to those of skill in the art.

3.6.7.1.1 DNA uptake

For mammalian cells that do not have cell walls, the calcium phosphate precipitation method for introduction of exogenous DNA [see, e. g., Graham et al. (1978) Virology 52:456-457; Wigler et al. (1979) Proc. Natl. Acad. Sci. U.S.A. 76:1373-1376; and Current Protocols in Molecular Biology, Vol. 1, Wiley Inter-Science, Supplement 14, Unit 9.1.1-9.1.9 (1990)] is often preferred. DNA uptake can be accomplished by DNA alone or in the presence of polyethylene glycol [PEG-mediated gene transfer], which is a fusion agent, or by any variations of such methods known to those of skill in the art [see, e.g., U.S. Pat. No. 4,684,611].

Lipid-mediated carrier systems are also among the preferred methods for introduction of DNA into cells [see, e.g., Teifel et al. (1995) Biotechniques 19:79-80; Albrecht et al. (1996) Ann. Hematol. 72:73-79; Holmen et al. (1995) In Vitro Cell

Dev. Biol. Anim. 31:347-351; Remy et al. (1994) Bioconjug. Chem. 5:647-654; Le Bolc h et al. (1995) Tetrahedron Lett. 36:6681-6684; Loeffler et al. (1993) Meth. Enzymol. 217:599-618]. Lipofection [see, e.g., Strauss (1996) Meth. Mol. Biol. 54:307-327] may also be used to introduce DNA into cells. This method is particularly well-suited for transfer of exogenous DNA into chicken cells (e.g., chicken blastodermal cells and primary chicken fibroblasts; see Brazolot et al. (1991) Mol. Repro. Dev. 30:304-312). In particular, DNA of interest can be introduced into chickens in operative linkage with promoters from genes, such as lysozyme and ovalbumin, that are expressed in the egg, thereby permitting expression of the heterologous DNA in the egg.

Additional methods useful in the direct transfer of DNA into cells include particle gun electrofusion [see, e.g., U.S. Pat. Nos. 4, 955,378, 4,923,814, 4,476,004, 4,906,576 and 4,441,972] and virion-mediated gene transfer.

A commonly used approach for gene transfer in land plants involves the direct introduction of purified DNA into protoplasts. The three basic methods for direct gene transfer into plant cells include: 1) polyethylene glycol [PEG]-mediated DNA uptake, 2) electroporation- mediated DNA uptake and 3) microinjection. In addition, plants may be transformed using ultrasound treatment [see, e.g., International PCT application publication No. WO 91/00358].

3.6.7.1.2 Electroporation

Electroporation involves providing high-voltage electrical pulses to a solution containing a mixture of protoplasts and foreign DNA to create reversible pores in the membranes of plant protoplasts as well as other cells. Electroporation is generally used for prokaryotes or other cells, such as plants that contain substantial cell-wall barriers. Methods for effecting electroporation are well known [see, e.g., U.S. Pat. Nos. 4,784,737, 5,501,967, 5,501,662, 5,019,034, 5,503,999; see, also Frommet al. 1985) Proc. Natl. Acad. Sci. U.S.A. 82:5824-5828].

For example, electroporation is often used for transformation of plants [see, e.g., Ag Biotechnology News 7:3 and 17 (September/October 1990)]. In this technique, plant protoplasts are electroporated in the presence of the DNA of interest that also includes a phenotypic marker. Electrical impulses of high field strength reversibly permeabilize biomembranes allowing the introduction of the plasmids. Electroporated plant protoplasts reform the cell wall, divide, and form plant callus. Transformed plant cells will be identified by virtue of the expressed phenotypic marker. The exogenous DNA may be added to the protoplasts in any form such as, for example, naked linear, circular or supercoiled DNA, DNA encapsulated in liposomes, DNA in spheroplasts, DNA in other plant protoplasts, DNA complexed with salts, and other methods.

3.6.7.1.3 Microcells

The chromosomes can be transferred by preparing microcells containing an artificial chromosome and then fusing with selected target cells. Methods for such preparation and fusion of microcells are well known [see, e.g., U.S. Pat. Nos. 5,240,840, 4, 806,476, 5,298,429, 5,396,767, Fournier (1981) Proc. Natl. Acad. Sci. U. S.A. 78:6349-6353; and Lambert et al. (1991) Proc. NatI. Acad. Sci. U.S. A. 88:5907-59]. Microcell fusion, using microcells that contain an artificial chromosome, is a particularly useful method for introduction of MACs into avian cells, such as DT40 chicken pre-B cells [for a description of DT40 cell fusion, see, e.g., Dieken et al. (1996) Nature Genet. 12:174-182].

3.6.7.2. Hosts

Suitable hosts include any host known to be useful for introduction and expression of heterologous DNA. Of particular interest herein, animal and plant cells and tissues, including, but not limited to insect cells and larvae, plants, and animals, particularly transgenic animals, and animal cells. Other hosts include, but are not limited to mammals, birds, particularly fowl such as chickens, reptiles, amphibians, insects, fish, arachnids, tobacco, tomato, wheat, monocots, dicots and algae, and any host into which introduction of heterologous DNA is desired. Such introduction can be effected using the MACs provided herein, or, if necessary by using the MACs provided herein to identify species-specific centromeres and/or functional chromosomal units and then using the resulting centromeres or chromosomal units as artificial chromosomes, or alternatively, using the methods exemplified herein for production of MACs to produce species-specific artificial chromosomes.

3.6.7.2.1 Introduction of DNA into embryos for production of transgenic animals and introduction of DNA into animal cells

Transgenic animals can be produced by introducing exogenous genetic material into a pronucleus of a mammalian zygote by microinjection [see, e.g., U.S. Pat. Nos. 4,873,191 and 5,354,674; see, also, International PCT application publication No. WO95/14769, which is based on U.S. application Ser. No.

08/159,084]. The zygote is capable of development into a mammal. The embryo or zygote is transplanted into a host female uterus and allowed to develop. Detailed protocols and examples are set forth below.

Transgenic chickens can be produced by injection of dispersed blastodermal cells from Stage X chicken embryos into recipient embryos at a similar stage of development [see e.g., Etches et al. (1993) Poultry Sci. 72:882-889; Petitte et al. (1990) Development 108:185-189]. Heterologous DNA is first introduced into the donor blastodermal cells using methods such as, for example, lipofection [see, e.g., Brazolot et al. (1991) Mol. Repro. Dev. 30:304-312] or microcell fusion [see, e.g., Dieken et al. 1996) Nature Genet. 12:174-182]. The transfected donor cells are then injected into recipient chicken embryos [see e.g., Carsience et al. (1993) Development 117: 669-675]. The recipient chicken embryos within the shell are candled and allowed to hatch to yield a germline chimeric chicken.

DNA can be introduced into animal cells using any known procedure, including, but not limited to: direct uptake, incubation with polyethylene glycol [PEG], microinjection, electroporation, lipofection, cell fusion, microcell fusion, particle bombardment, including microprojectile bombardment [see, eg., U.S. Pat. No. 5,470,708, which provides a method for transforming unattached mammalian cells via particle bombardment], and any other such method. For example, the transfer of plasmid DNA in liposomes directly to human cells in situ has been approved by the FDA for use in humans [see, eg., Nabel, et al. (1990) Science 249:1285-1288 and U.S. Pat. No. 5,461,032].

3.6.7.2.2 Introduction of heterologous DNA into plants.

Numerous methods for producing or developing transgenic plants are available to those of skill in the art. The method used is primarily a function of the species of plant. These methods include, but are not limited to: direct transfer of DNA by processes, such as PEG-induced DNA uptake, protoplast fusion, microinjection, electroporation, and microprojectile bombardment [see, e.g., Uchimiya et al. (1989) J. of Biotech. 12: 1-20 for a review of such procedures, see, also, e.g., U.S. Pat. Nos. 5,436,392 and 5,489,520 and many others]. For purposes herein, when introducing a MAC, microinjection, protoplast fusion and particle gun bombardment are preferred.

Plant species, including tobacco, rice, maize, rye, soybean, Brassica napus, cotton, lettuce, potato and tomato, have been used to produce transgenic plants.

Tobacco and other species, such as petunias, often serve as experimental models in which the methods have been developed and the genes first introduced and expressed.

DNA uptake can be accomplished by DNA alone or in the presence of PEG, which is a fusion agent, with plant protoplasts or by any variations of such methods known to those of skill in the art [see, e.g., U.S. Pat. No. 4,684,611 to Schilperoot et al.]. Electroporation, which involves high-voltage electrical pulses to a solution containing a mixture of protoplasts and foreign DNA to create reversible pores, has been used, for example, to successfully introduce foreign genes into rice and Brassica napus. Microinjection of DNA into plant cells, including cultured cells and cells in intact plant organs and embryoids in tissue culture and microprojectile bombardment [acceleration of small high density particles, which contain the DNA, to high velocity with a article gun apparatus, which forces the particles to penetrate plant cell walls and membranes] have also been used. All plant cells into which DNA can be introduced and that can be regenerated from the transformed cells can be used to produce transformed whole plants which contain the transferred artificial chromosome. The particular protocol and means for introduction of the DNA into the plant host may need to be adapted or refined to suit the particular plant species or cultivar.

3.6.7.2.3 Insect cells

Insects are useful hosts for introduction of artificial chromosomes for numerous reasons, including, but not limited to: (a) amplification of genes encoding useful proteins can be accomplished in the artificial chromosome to obtain higher protein yields in insect cells; (b) insect cells support required post-translational modifications, such as glycosylation and phosphorylation, that can be required for protein biological functioning; (c) insect cells do not support mammalian viruses, and, thus, eliminate the problem of cross- contamination of products with such infectious agents; (d) this technology circumvents traditional recombinant baculovirus systems for production of nutritional, industrial or medicinal proteins in insect cell systems; (e) the low temperature optimum for insect cell growth (28°C.) permits reduced energy cost of production; (f) serum-free growth medium for insect cells permits lower production costs; (g) artificial chromosome-containing cells can be stored indefinitely at low temperature; and (h) insect larvae will be biological factories for production of nutritional, medicinal or industrial proteins by microinjection of fertilized insect eggs [see, eq., Joy et al. (1991) Current Science 66:145-150, which provides a method for microinjecting heterologous DNA into Bombyx mori eggs].

Either MACs or insect-specific artificial chromosomes [BUGACs] will be used to introduce genes into insects. It appears that MACs will function in insects to direct expression of heterologous DNA contained thereon. For example, a MAC containing the B. mori actin gene promoter fused to the lacZ gene has been generated by transfection of EC3/7C5 cells with a plasmid containing the fusion gene. Subsequent fusion of the B. mori cells with the transfected EC3/7C5 cells that survived selection yielded a MAC-containing insect-mouse hybrid cell line in which, beta-galactosidase expression was detectable.

Insect host cells include, but are not limited to, hosts such as Spodoptera frugiperda [caterpillar], Aedes aegypti [mosquito], Aedes albopictus [mosquito], Drosphila melanogaster [fruitfly], Bombyx mori [silkworm], Manduca sexta [tomato horn worm] and Trichoplusia ni [cabbage looper]. Efforts have been directed toward propagation of insect cells in culture. Such efforts have focused on the fall armyworm, Spodoptera frugiperda. Cell lines have been developed also from other insects such as the cabbage looper, Trichoplusia ni and the silkworm, Bombyx mori. It has also been suggested that analogous cell lines can be created using the tomato hornworm, Manduca sexta. To introduce DNA into an insect, it should be introduced into the larvae, and allowed to proliferate, and then the hemolymph recovered from the larvae so that the proteins can be isolated therefrom.

The preferred method herein for introduction of artificial chromosomes into insect cells is microinjection [see, e.g., Tamura et al. (1991) Bio Ind. 8:26-31; Nikolaev et al. (1989) Mol. Biol. (Moscow) 23:1177-87; and methods exemplified and discussed erein].

3.6.8 Applications for and Uses of Artificial chromosomes

Artificial chromosomes provide convenient and useful vectors, and in some instances [e.g., in the case of very large heterologous genes] the only vectors, for introduction of heterologous genes into hosts. Virtually any gene of interest is amenable to introduction into a host via artificial chromosomes. Such genes include, but are not limited to, genes that encode receptors, cytokines, enzymes, proteases, hormones, growth factors, antibodies, tumor suppressor genes, therapeutic products and multigene pathways.

The artificial chromosomes provided herein will be used in methods of protein and gene product production, particularly using insects as host cells for production of such products, and in cellular (e.g., mammalian cell) production systems in which the rtificial chromomsomes (particularly MACs) provide a reliable, stable and efficient means for optimizing the biomanufacturing of important compounds for medicine and industry. They are also intended for use in methods of gene therapy, and in for production of transgenic plants and animals [discussed above and below].

3.6.8.1 Gene Therapy

Any nucleic acid encoding a therapeutic gene product or product of a multigene pathway may be introduced into a host animal, such as a human, or into a target cell line for introduction into an animal, for therapeutic purposes. Such therapeutic purposes include, genetic therapy to cure or to provide gene products that are missing or defective, to deliver agents, such as anti-tumor agents, to targeted cells or to an animal, and to provide gene products that will confer resistance or reduce susceptibility to a pathogen or ameliorate symptoms of a disease or disorder. The following are some exemplary genes and gene products. Such exemplification is not intended to be limiting.

3.6.8.1.1 Anti-HIV ribozymes

As exemplified below, DNA encoding anti-HIV ribozymes can be introduced and expressed in cells using MACs, including the euchromatin- based minichromosomes and the SATACs. These MACs can be used to make a transgenic mouse that expresses a ribozyme and, thus, serves as a model for testing the activity of such ribozymes or from which ribozyme- producing cell lines can be made. Also, introduction of a MAC that encodes an anti-HIV ribozyme into human cells will serve as treatment for HIV infection. Such systems further demonstrate the viability of using any disease-specific ribozyme to treat or ameliorate a particular disease.

3.6.8.1.2 Tumor Suppressor Genes

Tumor suppressor genes are genes that, in their wild-type alleles, express proteins that suppress abnormal cellular proliferation. When the gene coding for a tumor suppressor protein is mutated or deleted, the resulting mutant protein or the complete lack of tumor suppressor protein expression may result in a failure to correctly regulate cellular proliferation. Consequently, abnormal cellular proliferation may take place, particularly if there is already existing damage to the cellular regulatory mechanism. A number of well-studied human tumors and tumor cell lines have been shown to have missing or nonfunctional tumor suppressor genes.

Examples of tumor suppression genes include, but are not limited to, the retinoblastoma susceptibility gene or RB gene, the p53 gene, the gene that is deleted in colon carcinoma [i.e., the DCC gene] and the neurofibromatosis type 1 [NF-1] tumor suppressor gene [see, e.g., U.S. Pat. No. 5,496,731; Weinberg et al. (1991) 254:1138-1146]. Loss of function or inactivation of tumor suppressor genes may play a central role in the initiation and/or progression of a significant number of human cancers.

3.6.8.1.2.1 The p53 Gene

Somatic cell mutations of the p53 gene are said to be the most frequent of the gene mutations associated with human cancer [see, e.g., Weinberg et al. (1991) Science 254:1138-1146]. The normal or wild-type p53 gene is a negative regulator of cell growth, which, when damaged, favors cell transformation. The p53 expression product is found in the nucleus, where it may act in parallel or cooperatively with other gene products. Tumor cell lines in which p53 has been deleted have been successfully treated with wild-type p53 vector to reduce tumorigenicity [see, Baker et al. (1990) Science 249:912-915].

DNA encoding the p53 gene and plasmids containing this DNA are well known [see, e.g., U.S. Pat. No. 5,260,191; see, also Chen et al. (1990) Science 250:1576; Farrel et al. (1991) EMBO J. 10:2879-2887; plasmids containing the gene are available from the ATCC, and the sequence is in the GenBank Database, accession nos. X54156, X60020, M14695, M16494, K03199].

3.6.8.1.3 The CFTR gene

Cystic fibrosis [CF] is an autosomal recessive disease that affects epithelia of the airways, sweat glands, pancreas, and other organs. It is a lethal genetic disease associated with a defect in chloride ion transport, and is caused by mutations in the gene coding for the cystic fibrosis transmembrane conductance regulator [CFTR], a 1480 amino acid protein that has been associated with the expression of chloride conductance in a variety of eukaryotic cell types. Defects in CFTR destroy or reduce the ability of epithelial cells in the airways, sweat glands, pancreas and other tissues to transport chloride ions in response to cAMP-mediated agonists and impair activation of apical membrane channels by cAMP-dependent protein kinase A [PKA]. Given the high incidence and devastating nature of this disease, development of effective CF treatments is imperative.

The CFTR gene [~250 kb] [~600 kb] can be transferred into a MAC for use, for example, in gene therapy as follows. A CF-YAC [see Green et al. Science 250:94-98] may be modified to include a selectable marker, such as a gene encoding a protein that confers resistance to puromycin or hygromycin, and lambda-DNA for use in site-specific integration into a neo-minichromosome or a SATAC. Such a modified CF-YAC can be introduced into MAC-containing cells, such as EC3/7C5 or 19C5xHa4 cells, by fusion with yeast protoplasts harboring the modified CF-YAC or microinjection of yeast nuclei harboring the modified CF-YAC into the cells. Stable transformants are then selected on the basis of antibiotic resistance. These transformants will carry the modified CF-YAC within the MAC contained in the cells.

3.6.8.2 Animals, birds, fish and plants that are genetically altered to possess desired traits such as resistance to disease

Artificial chromosomes are ideally suited for preparing animals, including vertebrates and invertebrates, including birds and fish as well as mammals, that possess certain desired traits, such as, for example, disease resistance, resistance to harsh environmental conditions, altered growth patterns, and enhanced physical characteristics.

One example of the use of artificial chromosomes in generating disease-resistant organisms involves the preparation of multivalent vaccines. Such vaccines include genes encoding multiple antigens that can be carried in a MAC, or species-specific artificial chromosome, and either delivered to a host to induce immunity, or incorporated into embryos to produce transgenic animals and plants that are immune or less susceptible to certain diseases.

Disease-resistant animals and plants may also be prepared in which resistance or decreased susceptibility to disease is conferred by introduction into the host organism or embryo of artificial chromosomes containing DNA encoding gene products (e.g., ribozymes and proteins that are toxic to certain pathogens) that destroy or attenuate pathogens or limit access of pathogens to the host.

Animals and plants possessing desired traits that might, for example, enhance utility, processibility and commercial value of the organisms in areas such as the agricultural and ornamental plant industries may also be generated using artificial chromosomes in the same manner as described above for production of disease-resistant animals and plants. In such instances, the artificial chromosomes that are introduced into the organism or embryo contain DNA encoding gene products that serve to confer he desired trait in the organism.

Birds, particularly fowl such as chickens, fish and crustaceans will serve as model hosts for production of genetically altered organisms using artificial chromosomes.

3.6.8.3 Use of MACs and other artificial chromosomes for preparation and screening of libraries

Since large fragments of DNA can be incorporated into each artificial chromosome, the chromosomes are well-suited for use as cloning vehicles that can accommodate entire genomes in the preparation of genomic DNA libraries, which then can be readily screened. For example, MACs may be used to prepare a genomic DNA library useful in the identification and isolation of functional centromeric DNA from different species of organisms. In such applications, the MAC used to prepare a genomic DNA library from particular organism is one that is not functional in cells of that organism. That is, the MAC does not stably replicate, segregate or provide for expression of genes contained within it in cells of the organism. Preferably, the MACs contain an indicator gene (e.g., the lacZ gene encoding beta-galactosidase or genes encoding products that confer resistance to antibiotics such as neomycin, puromycin, hygromycin) linked to a promoter that is capable of promoting transcription of the indicator gene in cells of the organism. Fragments of genomic DNA from the organism are incorporated into the MACs, and the MACs are transferred to cells from the organism. Cells that contain MACs that have incorporated functional centromeres contained within the genomic DNA fragments are identified by detection of expression of the marker gene. For example, DNA encoding tree growth factors can be introduced into trees. Libraries can be prepared, introduce large fragments into chromosomes, and introduce them all into trees, thereby insuring expression.

3.6.8.4 Use of MACs and other artificial chromosomes for stable, high- level protein production

Cells containing the MACs and/or other artificial chromosomes provided herein are advantageously used for production of proteins, particularly several proteins from one cell line, such as multiple proteins involved in a biochemical pathway or multivalent vaccines. The genes encoding the proteins are introduced into the artificial chromosomes which are then introduced into cells. Alternatively, the heterologous gene(s) of interest are transferred into a production cell line that already contains artificial chromosomes in a manner that targets the gene(s) to the artificial chromosomes. The cells are cultured under conditions whereby the heterologous proteins are expressed. Because the proteins will be expressed at high levels in a stable permanent extra-genomic chromosomal system, selective conditions are not required.

Any transfectable cells capable of serving as recombinant hosts adaptable to continuous propagation in a cell culture system [see, e.g., McLean (1993) Trends In Biotech. 1 1 :232-238] are suitable for use in an artificial chromosome-based protein production system. Exemplary host cell lines include, but are not limited to, the following: Chinese hamster ovary (CHO) cells [see, e.g., Zang et al. (1995) Biotechnology 13:389-392], HEK 293, Ltk-, COS-7, DG44, and BHK cells. CHO cells are particularly preferred host cells. Selection of host cell lines for use in artificial chromosome-based protein production systems is within the skill of the art, but often will depend on a variety of factors, including the properties of the heterologous protein to be produced, potential toxicity of the protein in the host cell, any requirements for post-translational modification (e.g., glycosylation, amination, phosphorylation) of the protein, transcription factors available in the cells, the type of promoter element(s) being used to drive expression of the heterologous gene, whether production will be completely intracellular or the heterologous protein will preferably be secreted from the cell, and the types of processing enzymes in the cell.

The artificial chromosome-based system for heterologous protein production has many advantageous features. For example, as described above, because the heterologous DNA is located in an independent, extra- genomic artificial chromosome (as opposed to randomly inserted in an unknown area of the host cell genome or located as extrachromosomal element(s) providing only transient expression) it is stably maintained in an active transcription unit and is not subject to ejection via recombination or elimination during cell division. Accordingly, it is unnecessary to include a selection gene in the host cells and thus growth under selective conditions is also unnecessary. Furthermore, because the artificial chromosomes are capable of incorporating large segments of DNA, multiple copies of the heterologous gene and linked promoter element(s) can be retained in the chromosomes, thereby providing for high-level expression of the foreign protein(s). Alternatively, multiple copies of the gene can be linked to a single promoter element and several different genes may be linked in a fused polygene complex to a single promoter for expression of, for example, all the key proteins constituting a complete metabolic pathway [see, e.g., Beck von Bodman et al. (1995) Biotechnology 13:587-591]. Alternatively, multiple copies of a single gene can be operatively linked to a single promoter, or each or one or several copies may be linked to different promoters or multiple copies of the same promoter. Additionally, because artificial chromosomes have an almost unlimited capacity for integration and expression of foreign genes, they can be used not only for the expression of genes encoding end-products of interest, but also for the expression of genes associated with optimal maintenance and metabolic management of the host cell, e.g., genes encoding growth factors, as well as genes that may facilitate rapid synthesis of correct form of the desired heterologous protein product, e.g., genes encoding processing enzymes and transcription factors.

The MACS are suitable for expression of any proteins or peptides, including proteins and peptides that require in vivo posttranslational modification for their biological activity. Such proteins include, but are not limited to antibody fragments, full-length antibodies, and multimeric antibodies, tumor suppressor proteins, naturally occurring or artificial antibodies and enzymes, heat shock proteins, and others.

Thus, such cell-based "protein factories" employing MACs can generated using MACs constructed with multiple copies [theoretically an unlimited number or at least up to a number such that the resulting MAC is about up to the size of a genomic chromosome] of protein-encoding genes with appropriate promoters, or multiple genes driven by a single promoter, i.e., a fused gene complex [such as a complete metabolic pathway in plant expression system; see, e.g., Beck von Bodman (1995) Biotechnology 13:587-591]. Once such MAC is constructed, it can be transferred to a suitable cell culture system, such as a CHO cell line in protein-free culture medium [see, e.g., (1995) Biotechnology 13:389- 39] or other immortalized cell lines [see, e.g., (1993) TIBTECH 11:232- 238] where continuous production can be established.

The ability of MACs to provide for high-level expression of heterologous proteins in host cells is demonstrated, for example, by analysis of the H1D3 and G3D5 cell lines described herein and deposited with the ECACC. Northern blot analysis of mRNA obtained from these cells reveals that expression of the hygromycin-resistance and beta- galactosidase genes in the cells correlates with the amplicon number of the megachromosome(s) contained therein.

4. ENGINEERING APPROACHES

4.1.1 GENERAL CONSIDERATIONS

In one aspect, this invention applies the technical field of molecular genetics to evolve the genomes of cells and organisms to acquire new and improved properties.

Cells have a number of well-established uses in molecular biology. For example, cells are commonly used as hosts for manipulating DNA in processes such as transformation and recombination. Cells are also used for expression of recombinant proteins encoded by DNA transformed into the cells. Some types of cells are also used as progenitors for generation of transgenic animals and plants. Although all of these processes are now routine, in general, the genomes of the cells used in these processes have evolved little from the genomes of natural cells, and particularly not toward acquisition of new or improved properties for use in the above processes.

The traditional approach to artificial or forced molecular evolution focuses on optimization of an individual gene having a discrete and selectable phenotype. The strategy is to clone a gene, identify a discrete function for the gene and an assay by which it can be selected, mutate selected positions in the gene (e.g., by error-prone PCR or cassette mutagenesis) and select variants of the gene for improvement in the known function of the gene. A variant having improved function can then be expressed in a desired cell type. This approach has a number of limitations. First, it is only applicable to genes that have been isolated and functionally characterized. Second, the approach is usually only applicable to genes that have a discrete function. In other words, multiple genes that cooperatively confer a single phenotype cannot usually be optimized in this manner.

Probably, most genes do have cooper [illegible] explore a very limited number of the total number of permutations even for a single gene. For example, varying even ten positions in a protein with every possible amino acid would generate 20^{10} variants, which is more than can be accommodated by existing methods of transfection and screening.

In view of these limitations, the traditional approach is inadequate for improving cellular genomes in many useful properties. For example, to improve a cell's capacity to express a recombinant protein might require modification in any or all of a substantial number of genes, known and unknown, having roles in transcription, translation, posttranslational modification, secretion or proteolytic degradation, among others. Attempting individually to optimize even all the known genes having such functions would be a virtually impossible task, let alone optimizing hitherto unknown genes which may contribute to expression in manners not yet understood.

The present invention provides inter alia novel methods for evolving the genome of whole cells and organisms which overcome the difficulties and limitations of prior methods.

This ability to evolve genes artificially is of fundamental importance. For example, cells have a number of well-established uses in molecular biology, medicine and industrial processes. For example, cells are commonly used as hosts for manipulating DNA in processes such as transformation and recombination. Cells are used for expression of recombinant proteins encoded by DNA transformed/transfected or otherwise introduced into the cells. Some types of cells are used as progenitors for generation of transgenic animals and plants. The genomes of the cells used in these processes had evolved little from the genomes of natural cells, and particularly not toward acquisition of new or improved properties for use in the above processes.

Additional methods of recursively recombining nucleic acids in vivo and selecting resulting recombinants would be of use. The present invention provides a number of new and valuable methods and compositions for whole and partial genome evolution.

Metabolic engineering is the manipulation of intermediary metabolism through the use of both classical genetics and genetic engineering techniques. Cellular engineering is generally a more inclusive term referring to the modification of cellular properties. Cameron et al. (Applied Biochem. Biotech. 38:105-140 (1993)) provide a summary of equivalent terms to describe this type of engineering, including "metabolic engineering", which is most often used in the context of industrial microbiology and bioprocess engineering, "in vitro evolution" or "directed evolution", most often used in the context of environmental microbiology, "molecular breeding", most often used by Japanese researchers, "cellular engineering", which is used to describe modifications of bacteria, animal, and plant cells, "rational strain development", and "metabolic pathway evolution". In this application, the terms "metabolic engineering" and "cellular engineering" are used preferentially for clarity; the term "evolved" genes is used as discussed below.

Metabolic engineering can be divided into two basic categories: modification of genes endogenous to the host organism to alter metabolite flux and introduction of foreign genes into an organism. Such introduction can create new metabolic pathways leading to modified cell properties including but not limited to synthesis of known compounds not normally made by the host cell, production of novel compounds (e.g. polymers, antibiotics, etc.) and the ability to utilize new nutrient sources. Specific applications of metabolic engineering can include the production of specialty and novel chemicals, including antibiotics, extension of the range of substrates used for growth and product formation, the production of new catabolic activities in an organism for toxic chemical degradation, and modification of cell properties such as resistance to salt and other environmental factors.

Bailey (Science 252:1668-1674 (1991)) describes the application of metabolic engineering to the recruitment of heterologous genes for the improvement of a strain, with the caveat that such introduction can result in new compounds that may subsequently undergo further reactions, or that expression of a heterologous protein can result in proteolysis, improper folding, improper modification, or unsuitable intracellular location of the protein, or lack of access to required substrates. Bailey recommends careful configuration of a desired genetic change with minimal perturbation of the host. Liao (Curr. Opin. Biotech. 4:211-216 (1993)) reviews mathematical modeling and analysis of metabolic pathways, pointing out that in many cases the kinetic parameters of enzymes are unavailable or inaccurate.

Stephanopoulos et al. (Trends. Biotechnol. 11:392-396 (1993)) describe attempts to improve productivity of cellular systems or effect radical alteration of the flux through primary metabolic pathways as having difficulty in that control architectures at key branch points have evolved to resist flux changes. They conclude that identification and characterization of these metabolic nodes is a prerequisite to rational metabolic engineering. Similarly, Stephanopoulos (Curr. Opin. Biotech. 5:196-200 (1994)) concludes that rather than modifying the "rate limiting step" in metabolic engineering, it is necessary to systematically elucidate the control architecture of bioreaction networks. The present invention is generally directed to the evolution of new metabolic pathways and the enhancement of bioprocessing through a process herein termed recursive sequence recombination. Recursive sequence recombination entails performing iterative cycles of recombination and screening or selection to "evolve" individual genes, whole plasmids or viruses, multigene clusters, or even whole genomes (Stemmer, Bio/Technolog 13:549-553 (1995)). Such techniques do not require the extensive analysis and computation required by conventional methods for metabolic engineering. Recursive sequence recombination allows the recombination of large numbers of mutations in a minimum number of selection cycles, in contrast to traditional, pair wise recombination events.

Thus, because metabolic and cellular engineering can pose the particular problem of the interaction of many gene products and regulatory mechanisms, recursive sequence recombination (RSR) techniques provide particular advantages in that they provide recombination between mutations in any or all of these, thereby providing a very fast way of exploring the manner in which different combinations of mutations can affect a desired result, whether that result is increased yield of a metabolite, altered catalytic activity or substrate specificity of an enzyme or an entire metabolic pathway, or altered response of a cell to its environment.

4.1.2 THE EVOLUTIONARY IMPORTANCE OF RECOMBINATION

Strain improvement is the directed evolution of an organism to be more "fit" for a desired task. In nature, adaptation is facilitated by sexual recombination. Sexual recombination allows a population to exploit the genetic diversity within it, e.g., by consolidating useful mutations and discarding deleterious ones. In this way, adaptation and evolution can proceed in leaps. In the absence of a sexual cycle, members of a population must evolve independently by accumulating random mutations sequentially. Many useful mutations are lost while deleterious mutations can accumulate. Adaptation and evolution in this way proceeds slowly as compared to sexual evolution.

Asexual evolution is a slow and inefficient process.

Populations move as individuals rather than as groups. A diverse population is generated by the mutagenesis of a single parent resulting in a distribution of fit and unfit individuals. In the absence of a sexual cycle, each piece of genetic information of the surviving population remains in the individual mutants. Selection of the "fittest" results in many "fit" individuals being discarded along with the useful genetic information they carry. Asexual evolution proceeds one genetic event at a time and is thus limited by the intrinsic value of a single genetic event. Sexual evolution moves more quickly and efficiently. Mating within a population consolidates genetic information within the population and results in useful mutations being combined together. The combining of useful genetic information results in progeny that are much more fit than their parents. Sexual evolution thus proceeds much faster by multiple genetic events.

Years of plant and animal breeding has demonstrated the power of employing sexual recombination to effect the rapid evolution of complex genomes towards a particular task. This general principle is further demonstrated by using DNA stochastic &/or non-stochastic mutagenesis to recombine DNA molecules in vitro to accelerate the rate of directed molecular evolution. The strain improvement efforts of the fermentation industry rely on the directed evolution of microorganisms by sequential random mutagenesis. Incorporation of recombination into this iterative process greatly accelerates the strain improvement process, which in turn increases the profitability of current fermentation processes and facilitates the development of new products.

4.1.2.1 DNA STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS VS NATURAL RECOMBINATION

DNA stochastic &/or non-stochastic mutagenesis includes the recursive recombination of DNA sequences. A significant difference between DNA stochastic &/or non-stochastic mutagenesis and natural sexual recombination is that DNA stochastic &/or non-stochastic mutagenesis can produce DNA sequences originating from multiple parental sequences while sexual recombination produces DNA sequences originating from only two parental sequences.

The rate of evolution is in part limited by the number of useful mutations that a member of a population can accumulate between selection events.

In sequential random mutagenesis, useful mutations are accumulated one per selection event. Many useful mutations are discarded each cycle in favor of the best performer, and neutral or deleterious mutations which survive are as difficult to lose as they were to gain and thus accumulate. In sexual evolution pairwise recombination allows mutations from two different parents to segregate and recombine in different combinations. Useful mutations can accumulate and deleterious mutations can be lost. Poolwise recombination, such as that effected by DNA stochastic &/or non-stochastic mutagenesis, has the same advantages as pairwise recombination but allows mutations from many parents to consolidate into a single progeny. Thus poolwise recombination provides a means for increasing the number of useful mutations that can accumulate each selection event. One can plot the potential number of mutations an individual can accumulate by each of these processes. Recombination is exponentially superior to sequential random mutagenesis, and this advantage increases exponentially with the number of parents that can recombine. Sexual recombination is thus more conservative. In nature, the pairwise nature of sexual recombination may provide important stability within a population by impeding the large changes in DNA sequence that can result from poolwise recombination. For the purposes of directed evolution, however, poolwise recombination is more efficient.

The potential diversity that can be generated from a population is greater as a result of poolwise recombination as compared to that resulting from pairwise recombination. Further, poolwise recombination enables the combining of multiple beneficial mutations originating from multiple parental sequences. To demonstrate the importance of poolwise recombination vs pairwise recombination in the generation of molecular diversity consider the breeding of ten independent DNA sequences each containing only one unique mutation. There are $2^{10} = 1024$ different combinations of those ten mutations ranging from a single sequence having no mutations (the consensus) to that having all ten mutations. If this pool were recombined together by pairwise recombination, a population containing the consensus, the parents, and the 45 different combinations of any two of the mutations would result in 56 or ca. 5% of the possible 1024 mutant combinations. Alternatively, if the pool were recombined together in a poolwise fashion, all 1024 would be theoretically generated, resulting in an approximately 20 fold increase in library diversity. When looking for a unique solution to a problem in molecular evolution, the more complex the library, the more complex the possible solution. Indeed, the most fit member of a stochastic &/or non-stochastic mutagenized library often contains several mutations originating from several independent starting sequences.

4.1.2.2 DNA STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS PROVIDES RECURSIVE PAIRWISE RECOMBINATION

In vitro DNA stochastic &/or non-stochastic mutagenesis results in the efficient production of combinatorial genetic libraries by catalyzing the recombination of multiple DNA sequences. While the result of DNA stochastic &/or non-stochastic mutagenesis is a population representing the poolwise recombination of multiple sequences, the process does not rely on the recombination of multiple DNA sequences simultaneously, but rather on their recursive pairwise recombination. The assembly of complete genes from a mixed pool of small gene fragments requires multiple annealing and elongation cycles, the thermal cycles of the primerless PCR reaction. During each thermal cycle many pairs of fragments anneal and are extended to form a combinatorial population of larger chimeric DNA fragments. After the first cycle of stochastic &/or non-stochastic mutagenesis, chimeric fragments contain sequence originating from predominantly two different parent genes, with all possible pairs of "parental" sequence theoretically represented. This is similar to the result of a single sexual cycle within a population. During the second cycle, these chimeric fragments anneal with each other or with other small fragments, resulting in chimeras originating from up to four of the different starting sequences, again with all possible combinations of the four parental sequences theoretically represented. This second cycle is analogous to the entire population resulting from a single sexual cross, both parents and offspring, inbreeding.

Further cycles result in chimeras originating from 8, 16, 32, etc parental sequences and are analogous to further inbreedings of the preceding population. This could be considered similar to the diversity generated from a small population of birds that are isolated on an island, breeding with each other for many generations. The result mimics the outcome of "poolwise" recombination, but the path is via recursive pairwise recombination. For this reason, the DNA molecules generated from in vitro DNA stochastic &/or non-stochastic mutagenesis are not the "progeny" of the starting "parental" sequences, but rather the great, great great, great$_n$, (n = number of thermal cycles) grand progeny of the starting "ancestor" molecules.

4.1.3 DEFINITIONS

The term “cognate” refers to a gene sequence that is evolutionarily and functionally related between species. For example, in the human genome, the human CD4 gene is the cognate gene to the mouse CD4 gene, since the sequences and structures of these two genes indicate that they are homologous and that both genes encode a protein which functions in signaling T-cell activation through MHC class II-restricted antigen recognition. Screening is, in general, a two-step process in which one first determines which cells do and do not express a screening marker or phenotype (or a selected level of marker or phenotype), and then physically separates the cells having the desired property. Selection is a form of screening in which identification and physical separation are achieved simultaneously by expression of a selection marker, which, in some genetic circumstances, allows cells expressing the marker to survive while other cells die (or vice versa). Screening markers include luciferase, P-galactosidase, and green fluorescent protein. Selection markers include drug and toxin resistance genes.

An exogenous DNA segment is one foreign (or heterologous) to the cell or homologous to the cell but in a position within the host cell nucleic acid in which the element is not ordinarily found. Exogenous DNA segments can be expressed to yield exogenous polypeptides.

The term "gene" is used broadly to refer to any segment of DNA associated with a biological function. Thus, genes include coding sequences and/or the regulatory sequences required for their expression. Genes also include nonexpressed DNA segments that, for example, form recognition sequences for other proteins.

The terms "identical" or "percent identity," in the context of two or more nucleic acids or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same, when compared and aligned for maximum correspondence, as measured using one of the following sequence comparison algorithms or by visual inspection.

The phrase "substantially identical," in the context of two nucleic acids or polypeptides, refers to two or more sequences or subsequences that have at least 60%, preferably 80%, most preferably 90-95% nucleotide or amino acid residue identity, when compared and aligned for maximum correspondence, as measured using one of the following sequence comparison algorithms or by visual inspection. Preferably, the substantial identity exists over a region of the sequences that is at least about 50 residues in length, more preferably over a region of at least about 100 residues, and most preferably the sequences are substantially identical over at least about 150 residues. In a most preferred embodiment, the sequences are substantially identical over the entire length of the coding regions.

For sequence comparison, typically one sequence acts as a reference sequence, to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters.

Optimal alignment of sequences for comparison can be conducted, e.g., by the local homology algorithm of Smith & Waterman, Adv. Appl Math. 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, J Mol Biot 48:443 (1970), by the search for similarity method of Pearson & Lipman, Proc. Natl. Acad. Sci. USA 85:2444 (1988), by computerized implementations of algorithms GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package Release 7.0, Genetics Computer Group, 575 Science Dr., Madison, W1.

Another example of a useful alignment algorithm is PILEUP. PILEUP creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments to show relationship and percent sequence identity. It also plots a tree or dendogram showing the clustering relationships used to create the alignment. PILEUP uses a simplification of the progressive alignment method of Feng &

Doolittle, J Mol. Evol. 35:351 - 360 (1987). The method used is similar to the method described by Higgins & Sharp, CABIOS 5:151-153 (1989). The program can align up to 300 sequences, each of a maximum length of 5,000 nucleotides or amino acids. The multiple alignment procedure begins with the pairwise alignment of the two most similar sequences, producing a cluster of two aligned sequences. This cluster is then aligned to the next most related sequence or cluster of aligned sequences. Two clusters of sequences are aligned by a simple extension of the pairwise alignment of two individual sequences. The final alignment is achieved by a series of progressive, pairwise alignments. The program is run by designating specific sequences and their amino acid or nucleotide coordinates for regions of sequence comparison and by designating the program parameters. For example, a reference sequence can be compared to other test sequences to determine the percent sequence identity relationship using the following parameters: default gap weight (3. 00), default gap length weight (0. 10), and weighted end gaps.

Another example of algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul et al., J Mol. Biol. 215:403-410 (1990). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul et al, supra). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative- scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a word length (W) of 11, an expectation (E) of 10, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a word length (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (see Henikoff & Henikoff, Proc. Natl. Acad. Sci. USA 89:10915 (1989)).

In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (see, e.g., Karlin & Altschul, Proc. Natl. Acad. Sci. USA 90:5873-5787 (1993)). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0. 1, more preferably less than about 0. 0 1, and most preferably less than about 0.001.

A further indication that two nucleic acid sequences or polypeptides are substantially identical is that the polypeptide encoded by the first nucleic acid is immunologically cross reactive with the polypeptide encoded by the second nucleic acid, as described below. Thus, a polypeptide is typically substantially identical to a second polypeptide, for example, where the two peptides differ only by conservative substitutions.

Another indication that two nucleic acid sequences are substantially identical is that the two molecules hybridize to each other under stringent conditions.

The term "naturally-occurring" is used to describe an object that can be found in nature. For example, a polypeptide or polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a source in nature and which has not been intentionally modified by man in the laboratory is naturally- occurring. Generally, the term naturally-occurring refers to an object as present in a non-pathological (undiseased) individual, such as would be typical for the species.

Asexual recombination is recombination occurring without the fusion of gametes to form a zygote.

A "mismatch repair deficient strain" can include any mutants in any organism impaired in the functions of mismatch repair. These include mutant gene products of mutS, mutT, mutH, mutL, ovrD, dcm, vsr, umuC, umuD, sbcB, recJ, etc. The impairment is achieved by genetic mutation, allelic replacement, selective inhibition by an added reagent such as a small compound or an expressed antisense RNA, or other techniques. Impairment can be of the genes noted, or of homologous genes in any organism.

4.2. STRATEGIES

4.2.1. EVOLVING A CELL TO ACQUIRE A DESIRED FUNCTION

4.2.1.1. DESIRED FUNCTION IS SECRETION OF A PROTEIN

Optionally, the desired function is secretion of a protein, and the plurality of cells further comprises a construct encoding the protein. The protein is optionally inactive unless secreted, and further modified cells are optionally selected for protein function.

Optionally, the protein is toxic to the plurality of cells, unless secreted. In this case, the modified or further modified cells which evolve toward acquisition of the desired function are screened by propagating the cells and recovering surviving cells.

4.2.1.2. DESIRED FUNCTION IS ENHANCED RECOMBINATION

In some methods, the desired function is enhanced recombination. In such methods, the library of fragments sometimes comprises a cluster of genes collectively conferring recombination capacity. Screening can be achieved using cells carrying a gene encoding a marker whose expression is prevented by a mutation removable by recombination. The cells are screened by their expression of the marker resulting from removal of the mutation by recombination.

4.2.13. DESIRED FUNCTION IS IMPROVED RESISTANCE IN PLANT CELLS

In some methods, the plurality of cells are plant cells and the desired property is improved resistance to a chemical or microbe. The modified or further modified cells (or whole plants) are exposed to the chemical or microbe and modified or further modified cells having evolved toward the acquisition of the desired function are selected by their capacity to survive the exposure.

4.2.1.4. DESIRED FUNCTION IS PREDICTIONG EFFICACY OF A DRUG

4.2.1.4.1. A DRUG TREATING A VIRAL INFECTION

The invention further provides methods of predicting efficacy of a drug in treating a viral infection. Such methods entail recombining a nucleic acid segment from a virus, whose infection is inhibited by a drug, with at least a second nucleic acid segment from the virus, the second nucleic acid segment differing from the first nucleic acid segment in at least two nucleotides, to produce a library of recombinant nucleic acid segments. Host cells are then contacted with a collection of viruses having genomes including the recombinant nucleic acid segments in a media containing the drug, and progeny viruses resulting from infection of the host cells are collected.

A recombinant DNA segment from a first progeny virus recombines with at least a recombinant DNA segment from a second progeny virus to produce a further library of recombinant nucleic acid segments. Host cells are contacted with a collection of viruses having genomes including the further library or recombinant nucleic acid segments, in media containing the drug, and further progeny viruses are produced by the host cells. The recombination and selection steps are repeated, as desired, until a further progeny virus has acquired a desired degree of resistance to the drug, whereby the degree of resistance acquired and the number of repetitions needed to acquire it provide a measure of the efficacy of the drug in treating the virus. Viruses are optionally adapted to grow on particular cell lines.

4.2.1.4.2. A DRUG TREATING INFECTION BY A PATHOGENIC MICROORGANISM

The invention further provides methods of predicting efficacy of a drug in treating an infection by a pathogenic microorganism. These methods entail delivering a library of DNA fragments into a plurality of microorganism cells, at least some of which undergo recombination with segments in the genome of the cells to produce modified microorganism cells. Modified microorganisms are propagated in a media containing the drug, and surviving microorganisms are recovered. DNA from surviving microorganisms is recombined with a further library of DNA fragments at least some of which undergo recombination with cognate segments in the DNA from the surviving microorganisms to produce further modified microorganisms cells. Further modified microorganisms are propagated in media containing the drug, and further surviving microorganisms are collected. The recombination and selection steps are repeated as needed, until a further surviving microorganism has acquired a desired degree of resistance to the drug. The degree of resistance acquired and the number of repetitions needed to acquire it provide a measure of the efficacy of the drug in killing the pathogenic microorganism.

4.2.1.3. METHOD

4.2.1.3.1 MODIFY OR RECOMBINE CELLS

In one aspect, the invention provides methods of evolving a cell to acquire a desired function. Such methods entail, e.g., introducing a library of DNA fragments into a plurality of cells, whereby at least one of the fragments undergoes recombination with a segment in the genome or an episome of the cells to produce modified cells. Optionally, these modified cells are bred to increase the diversity of the resulting recombined cellular population. The modified cells, or the recombined cellular population are then screened for modified or recombined cells that have evolved toward acquisition of the desired function. DNA from the modified cells that have evolved toward the desired function is then optionally recombined with a further library of DNA fragments, at least one of which undergoes recombination with a segment in the genome or the episome of the modified cells to produce further modified cells. The further modified cells are then screened for further modified cells that have further evolved toward acquisition of the desired function. Steps of recombination and screening/selection are repeated as required until the further modified cells have acquired the desired function. In one preferred embodiment, modified cells are recursively recombined to increase diversity of the cells prior to performing any selection steps on any resulting cells.

4.2.1.3.2 COAT WITH RecA

In some methods, the library or further library of DNA fragments is coated with recA protein to stimulate recombination with the segment of the genome. The library of fragments is optionally denatured to produce single-stranded DNA, which are annealed to produce duplexes, some of which contain mismatches at points of variation in the fragments. Duplexes containing mismatches are optionally selected by affinity chromatography to immobilized MutS.

4.2.1.3.3 PERFORM IN VIVO RECOMBINATION

The invention further provides methods for performing in vivo recombination. At least first and second segments from at least one gene are introduced into a cell, the segments differing from each other in at least two nucleotides, whereby the segments recombine to produce a library of chimeric genes. A chimeric gene is selected from the library having acquired a desired function.

The invention further provides methods of evolving a cell to acquire a desired function. These methods entail providing a populating of different cells. The cells are cultured under conditions whereby DNA is exchanged between cells, forming cells with hybrid genomes. The cells are then screened or selected for cells that have evolved toward acquisition of a desired property. The DNA exchange and screening/selecting steps are repeated, as needed, with the screened/selected cells from one cycle forming the population of different cells in the next cycle, until a cell has acquired the desired property.

Mechanisms of DNA exchange include conjugation, phage-mediated transduction, liposome delivery, protoplast fusion, and sexual recombination of the cells. Optionally, a library of DNA fragments can be transformed or electroporated into the cells.

4.2.1.3.4 PROTOPLAST-MEDIATED EXCHANGE

As noted, some methods of evolving a cell to acquire a desired property are effected by protoplast-mediated exchange of DNA between cells. Such methods entail forming protoplasts of a population of different cells. The protoplasts are then fused to form hybrid protoplasts, in which genomes from the protoplasts recombine to form hybrid genomes. The hybrid protoplasts are incubated under conditions promoting regeneration of cells. The regenerated cells can be recombined one or more times (i.e., via protoplasting or any other method than combines genomes of cells) to increase the diversity of any resulting cells. Preferably, regenerated cells are recombined several times, e.g., by protoplast fusion to generate a diverse population of cells. The next step is to select or screen to isolate regenerated cells that have evolved toward acquisition of the desired property. DNA exchange and selection/screening steps are repeated, as needed, with regenerated cells in one cycle being used to form protoplasts in the next cycle until the regenerated cells have acquired the desired property. Industrial microorganisms are a preferred class of organisms for conducting the above methods. Some methods further comprise a step of selecting or screening for fused protoplasts free from unfused protoplasts of parental cells. Some methods further comprise a step of selecting or screening for fused protoplasts with hybrid genomes free from cells with parental genomes. In some methods, protoplasts are provided by treating individual cells, mycelia or spores with an enzyme that degrades cell walls. In some methods, the strain is a mutant that is lacking capacity for intact cell wall synthesis, and protoplasts form spontaneously. In some methods, protoplasts are formed by treating growing cells with an inhibitor of cell wall formation to generate protoplasts. In some methods, the desired property is expression and/or secretion of a protein or secondary metabolite, such as an industrial enzyme, a therapeutic protein, a primary metabolite such as lactic acid or ethanol, or a secondary metabolite such as erythromycin cyclosporin A or taxol. In other methods it is the ability of the cell to convert compounds provided to the cell to different compounds. In yet other methods, the desired property is capacity for meiosis. In some methods, the desired property is compatibility to form a heterokaryon with another strain.

The invention further provides methods of evolving a cell toward acquisition of a desired property. These methods entail providing a population of different cells. DNA is isolated from a first subpopulation of the different cells and encapsulated in liposomes. Protoplasts are formed from a second subpopulation of the different cells. Liposomes are fused with the protoplasts, whereby DNA from the liposomes is taken up by the protoplasts and recombines with the genomes of the protoplasts. The protoplasts are incubated under regenerating conditions. Regenerating or regenerated cells are then selected or screened for evolution toward the desired property.

4.2.1.3.4 REITERATIVE POOLING AND BREEDING OF HIGHER ORGANISMS

The method also provides methods of reiterative pooling and breeding of higher organisms. In the methods, a library of diverse multicellular organisms are produced (e.g., plants, animals or the like). A pool of male gametes is provided along with a pool of female gametes. At least one of the male pool or the female pool comprises a plurality of different gametes derived from different strains of a species or different species. The male gametes are used to fertilize the female gametes. At least a portion of the resulting fertilized gametes grow into reproductively viable organisms. These reproductively viable organisms are crossed (e.g., by pairwise pooling and joining of the male and female gametes as before) to produce a library of diverse organisms. The library is then selected for a desired trait or property.

The library of diverse organisms can comprise a plurality of plants such as Gramineae, Fetucoideae, Poacoideae, Agrostis, Phleum, Dactylis, Sorgum, Setaria, Zea, Oryza, Triticum, Secale, Avena, Hordeum, Saccharum, Poa, Festuca, Stenotaphrum, Cynodon, Coix, Olyreae, Phareae, Compositae or Leguminosae. For example, the plants can be e.g., corn, rice, wheat, rye, oats, barley, pea, beans, lentil, peanut, yam bean, cowpeas, velvet beans, soybean, clover, alfalfa, lupine, vetch, lotus, sweet clover, wisteria, sweet pea, sorghum, millet, sunflower, canola or the like.

Similarly, the library of diverse organisms can include a plurality of animals such as non-human mammals, fish, insects, or the like.

Optionally, a plurality of selected library members can be crossed by pooling gametes from the selected members and repeatedly crossing any resulting additional reproductively viable organisms to produce a second library of diverse organisms (e.g., by split pair wise pooling and rejoining of the male and female gametes). Here again, the second library can be selected for a desired trait or property, with the resulting selected members forming the basis for additional poolwise breeding and selection. A feature of the invention is the libraries made by these (or any preceding) method.

4.3. ORIGIN OF CELLS

4.3.1 EMBRYONIC CELLS OF AN ANIMAL

In some methods, the plurality of cells are embryonic cells of an animal, and the method further comprises propagating the transformed cells to transgenic animals. The plurality of cells can be a plurality of industrial microorganisms that are enriched for microorganisms which are tolerant to desired process conditions (heat, light, radiation, selected pFL presence of detergents or other denaturants, presence of alcohols or other organic molecules, etc.).

4.3.2 ARTIFICIAL CHROMOSOMES

The invention further provides methods of evolving a cell toward acquisition of a desired property using artificial chromosomes. Such methods entail introducing a DNA fragment library cloned into an artificial chromosome into a population of cells. The cells are then cultured under conditions whereby sexual recombination occurs between the cells, and DNA fragments cloned into the artificial chromosome recombines by homologous recombination with corresponding segments of endogenous chromosomes of the populations of cells, and endogenous chromosomes recombine with each other. Cells can also be recombined via conjugation. Any resulting cells can be recombined via any method noted herein, as many times as desired, to generate a desired level of diversity in the resulting recombinant cells. In any case, after generating a diverse library of cells, the cells that have evolved toward acquisition of the desired property are screened and/or selected for a desired property. The method is then repeated with cells that have evolved toward the desired property in one cycle forming the population of different cells in the next cycle. Here again, multiple cycles of in vivo recombination are optionally performed prior to any additional selection or screening steps.

The invention further provides methods of evolving a DNA segment cloned into an artificial chromosome for acquisition of a desired property. These methods entail providing a library of variants of the segment, each variant cloned into separate copies of an artificial chromosome. The copies of the artificial chromosome are introduced into a population of cells. The cells are cultured under conditions whereby sexual recombination occurs between cells and homologous recombination occurs between copies of the artificial chromosome bearing the variants. Variants are then screened or selected for evolution toward acquisition of the desired property.

The invention further provides hyperrecombinogenic recA proteins.

4.4. METHOD TO ACQUIRE A BIOCATALYTIC ACTIVITY

One aspect of the invention is a method of evolving a biocatalytic activity of a cell, comprising:

(a) recombining at least a first and second DNA segment from at least one gene conferring ability to catalyze a reaction of interest, the segments differing from each other in at least two nucleotides, to produce a library of recombinant genes;

(b) screening at least one recombinant gene from the library that confers enhanced ability to catalyze the reaction of interest by the cell relative to a wild type form of the gene;

(c) recombining at least a segment from at least one recombinant gene with a further DNA segment from at least one gene, the same or different from the first and second segments, to produce a further library of recombinant genes;

(d) screening at least one further recombinant gene from the further library of recombinant genes that confers enhanced ability to catalyze the reaction of interest in the cell relative to a previous recombinant gene;

(e) repeating (c) and (d), as necessary, until the further recombinant gene confers a desired level of enhanced ability to catalyze the reaction of interest by the cell.

4.4.1. METHOD TO EVOLVE A GENE TO CATALYZE A RXN OF INTEREST

Another aspect of the invention is a method of evolving a gene to confer ability to catalyze a reaction of interest, the method comprising:

(1) recombining at least first and second DNA segments from at least one gene conferring ability to catalyze a reaction of interest, the segments differing from each other in at least two nucleotides, to produce a library of recombinant genes;

(2) screening at least one recombinant gene from the library that confers enhanced ability to catalyze a reaction of interest relative to a wild type form of the gene;

(3) recombining at least a segment from the at least one recombinant gene with a further DNA segment from the at least one gene, the same or different from the first and second segments, to produce a further library of recombinant genes;

(4) screening at least one further recombinant gene from the further library of recombinant genes that confers enhanced ability to catalyze a reaction of interest relative to a previous recombinant gene;

(5) repeating (3) and (4), as necessary, until the further recombinant gene confers a desired level of enhanced ability to catalyze a reaction of interest.

4.4.2. METHOD TO GENERATE A NEW BIOCATALYTIC ACTIVITY IN A CELL

A further aspect of the invention is a method of generating a new biocatalytic activity in a cell, comprising:

(1) recombining at least first and second DNA segments from at least one gene conferring ability to catalyze a first reaction related to a second reaction of interest, the segments differing from each other in at least two nucleotides, to produce a library of recombinant genes;

(2) screening at least one recombinant gene from the library that confers a new ability

to catalyze the second reaction of interest;

(3) recombining at least a segment from at least one recombinant gene with a further DNA segment from the at least one gene, the same or different from the first and second segments, to produce a further library of recombinant genes;

(4) screening at least one further recombinant gene from the further library of recombinant genes that confers enhanced ability to catalyze the second reaction of interest in the cell relative to a previous recombinant gene;

(5) repeating (3) and (4), as necessary, until the further recombinant gene confers a desired level of enhanced ability to catalyze the second reaction of interest in the cell.

4.4.3. METHOD TO MODIFY A METABOLIC PATHWAY EVOLVED BY RECURSIVE SEQUENCE RECOMBINATION

Another aspect of the invention is a modified form of a cell, wherein the modification comprises a metabolic pathway evolved by recursive sequence recombination.

A further aspect of the invention is a method of optimizing expression of a gene product, the method comprising:

(1) recombining at least first and second DNA segments from at least one gene conferring ability to produce the gene product, the segments differing from each other in at least two nucleotides, to produce a library of recombinant genes;

(2) screening at least one recombinant gene from the library that confers optimized expression of the gene product relative to a wild type form of the gene;

(3) recombining at least a segment from the at least one recombinant gene with a further DNA segment from the at least one gene, the same or different from the first and second segments, to produce a further library of recombinant genes;

(4) screening at least one further recombinant gene from the further library of recombinant genes that confers optimized ability to produce the gene product relative to a previous recombinant gene;

(5) repeating (3) and (4), as necessary, until the further recombinant gene confers a

desired level of optimized ability to express the gene product.

4.4.4. METHOD TO EVOLVE A BIOSENSOR

A further aspect of the invention is a method of evolving a biosensor for a compound A of interest, the method comprising:

(1) recombining at least first and second DNA segments from at least one gene conferring ability to detect a related compound B, the segments differing from each other in at least two nucleotides, to produce a library of recombinant genes;

(2) screening at least one recombinant gene from the library that confers optimized ability to detect compound A relative to a wild type form of the gene;

(3) recombining at least a segment from the at least one recombinant gene with a further DNA segment from the at least one gene, the same or different from the first and second segments, to produce a further library of recombinant genes;

(4) screening at least one further recombinant gene from the further library of recombinant genes that confers optimized ability to detect compound A relative to a previous recombinant gene;

(5) repeating (3) and (4), as necessary, until the further recombinant gene confers a desired level of optimized ability to detect compound A.

4.5. FERMENTATION OF MICRO-ORGANISMS

The fermentation of microorganisms for the production of natural products is the oldest and most sophisticated application of biocatalysis. Industrial microorganisms effect the multistep conversion of renewable feedstocks to high value chemical products in a single reactor and in so doing catalyze a multi-billion dollar industry. Fermentation products range from fine and commodity chemicals such as ethanol, lactic acid, amino acids and vitamins, to high value small molecule pharmaceuticals, protein pharmaceuticals, and industrial enzymes. (See, e.g., McCoy (1998) C&EN 13-

19) for an introduction to biocatalysis.

The methods herein allow biocatalysts to be improved at a faster pace than conventional methods. Whole genome stochastic &/or non-stochastic mutagenesis can at least double the rate of strain improvement for microorganisms used in fermentation as compared to traditional methods. This provides for a relative decrease in the cost of fermentation processes. New products can enter the market sooner, producers can increase profits as well as market share, and consumers gain access to more products of higher quality and at lower prices. Further, increased efficiency of production processes translates to less waste production and more frugal use of resources. Whole genome stochastic &/or non-stochastic mutagenesis provides a means of accumulating multiple useful mutation per cycle and thus eliminate the inherent limitation of current strain improvement programs (SIPs).

One key to SIP is having an assay that can be dependably used to identify a few mutants out of thousands that have subtle increases in product yield. The limiting factor in many assay formats is the uniformity of cell growth. This variation is the source of baseline variability in subsequent assays. Inoculum size and culture environment (temperature/humidity) are sources of cell growth variation. Automation of all aspects of establishing initial cultures and state-of-the-art temperature and humidity controlled incubators are useful in reducing variability.

Mutant cells or spores are separated on solid media to produce individual sporulating colonies. Using an automated colony picker (Q-bot, Genefix, U.K.), colonies are identified, picked, and 10,000 different mutants inoculated into 96 well microtitre dishes containing two 3 mm. glass balls/well. The Q-bot does not pick an entire colony but rather inserts a pin through the center of the colony and exits with a small sampling of cells (or mycelia) and spores. The time the pin is in the colony, the number of dips to inoculate the culture medium, and the time the pin is in that medium each effect inoculum size, and each can be controlled and optimized. The uniform process of the Q-bot decreases human handling error and increases the rate of establishing cultures (roughly 10,000/4 hours). These cultures are then shaken in a temperature and humidity controlled incubator. The glass balls act to promote uniform aeration of cells and the dispersal of mycelial fragments similar to the blades of a fermenter.

1. Prescreen The ability to detect a subtle increase in the performance of a mutant over that of a parent strain relies on the sensitivity of the assay. The chance of finding the organisms having an improvement is increased by the number of individual mutants that can be screened by the assay. To increase the chances of identifying a pool of sufficient size a prescreen that increases the number of mutants processed by 10-fold can be used. The goal of the primary screen will be to quickly identify mutants having equal or better product titres than the parent strain(s) and to move only these mutants forward to liquid cell culture. The primary screen is an agar plate screen is analyzed by the Q-bot colony picker. Although assays can be fundamentally different, many result, e.g., in the production of colony halos. For example, antibiotic production is assayed on plates using an overlay of a sensitive indicator strain, such as B. subtilis. Antibiotic production is typically assayed as a zone of clearing (inhibited growth of the indicator organism) around the producing organism. Similarly, enzyme production can be assayed on plates containing the enzyme substrate, with activity being detected as a zone of substrate modification around the producing colony. Product titre is correlated with the ratio of halo area to colony area.

The Q-bot or other automated system is instructed to only pick colonies having a halo ratio in the top 10% of the population i.e. 10,000 mutants from the 100,000 entering the plate prescreen. This increases the number of improved clones in the secondary assay and eliminates the wasted effort of screening knock-out and low producers. This improves the "hit rate" of the secondary assay.

4.6. EXPERIMENTAL APPLICATIONS

4.6.1 STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS

4.6.1.1 GENERAL TECHNIQUES

4.6.1.1.1 STARTING MATERIALS

Thus, a general method for recursive sequence recombination for the embodiments herein is to begin with a gene encoding an enzyme or enzyme subunit and to evolve that gene either for ability to act on a new substrate, or for enhanced catalytic properties with an old substrate, either alone or in combination with other genes in a multistep pathway. The term "gene" is used herein broadly to refer to any segment or sequence of DNA associated with a biological function. Genes can be obtained from a variety of sources, including cloning from a source of interest or synthesizing from known or predicted sequence information, and may include sequences designed to have desired parameters. The ability to use a new substrate can be assayed in some instances by the ability to grow on a substrate as a nutrient source. In other circumstances such ability can be assayed by decreased toxicity of a substrate for a host cell, hence allowing the host to grow in the presence of that substrate. Biosynthesis of new compounds, such as antibiotics, can be assayed similarly by growth of an indicator organism in the presence of the host expressing the evolved genes. For example, when an indicator organism used in an overlay of the host expressing the evolved gene(s), wherein the indicator organism is sensitive or expected to be sensitive to the desired antibiotic, growth of the indicator organism would be inhibited in a zone around the host cell or colony expressing the evolved gene(s).

Another method of identifying new compounds is the use of standard analytical techniques such as mass spectroscopy, nuclear magnetic resonance, high performance liquid chromatography, etc. Recombinant microorganisms can be pooled and extracts or media supernatants assayed from these pools. Any positive pool can then be subdivided and the procedure repeated until the single positive is identified ("sib-selection").

In some instances, the starting material for recursive sequence recombination is a discrete gene, cluster of genes, or family of genes known or thought to be associated with metabolism of a particular class of substrates. One of the advantages of the instant invention is that structural information is not required to estimate which parts of a sequence should be mutated to produce a functional hybrid enzyme.

In some embodiments of the invention, an initial screening of enzyme activities in a particular assay can be useful in identifying candidate enzymes as starting materials. For example, high throughput screening can be used to screen enzymes for dioxygenase-type activities using aromatic acids as substrates. Dioxygenases typically transform indole-2- carboxylate and indole-3-carboxylate to colored products, including indigo (Eaton et. al. J. Bacteriol. 177:6983-6988 (1995)). DNA encoding enzymes that give some activity in the initial assay can then be recombined by the recursive techniques of the invention and rescreened. The use of such initial screening for candidate enzymes against a desired target molecule or analog of the target molecule can be especially useful to generate enzymes that catalyze reactions of interest such as catabolism of man-made pollutants.

This type of high throughput screening can also be used during each round of recursive sequence recombination to identify mutants that possess the highest level of the desired activity. For example, penicillin G acylases have been isolated by looking for clones that allow a leucine auxotroph to hydrolyse penicillin G analogue phenylacetyl-L-leucine, thereby producing leucine and allowing cell growth (Martin, L. et al., FEMS Microbiology Lett. 125:287-292 (1995)). Positives from this selection are then screened by a more labour-intensive method for ability to hydrolyse penicillin G.

This same selection on phenylacetyl-L-leucine can be used when evolving penicillin G acylase for greater activity by recursive sequence recombination. After each round of recombination the library of acylase genes is transformed into a leucine auxotroph. Those that grow fastest are picked as probably having the most active acylase. The acylases are then be tested against the real substrate, penicillin G, by a more laborious screen such as HPLC. Thus, even if there is no convenient high throughput screen for an enzyme or a metabolic pathway, it is often possible to find a rapid detection method that can approximately measure the desired phenotype, thereby reducing the numbers of colonies that must be screened more accurately.

The starting material can also be a segment of such a gene or cluster that is recombined in isolation of its surrounding DNA, but is relinked to its surrounding DNA before screening/selection of recombination products. In other instances, the starting material for recombination is a larger segment of DNA that includes a coding sequence or other locus associated with metabolism of a particular substrate at an unknown location. For example, the starting material can be a chromosome, episome, YAC, cosmid, or phage P1 clone. In still other instances, the starting material is the whole genome of an organism that is known to have desirable metabolic properties, but for which no information localizing the genes associated with these characteristics is available.

In general any type of cells can be used as a recipient of evolved genes. Cells of particular interest include many bacterial cell types, both gram-negative and gram-positive, such as Rhodococcus, Streptomycetes, Actinomycetes, Corynebacteria, Penicillium, Bacillus, Escherichia coli, Pseudomonas, Salmonella, and Erwinia. Cells of interest also include eukaryotic cells, particularly mammalian cells (e.g., mouse, hamster, primate, human), both cell lines and primary cultures. Such cells include stem cells, including embryonic stem cells, zygotes, fibroblasts, lymphocytes, Chinese hamster ovary (CHO), mouse fibroblasts (NIHM), kidney, liver, muscle, and skin cells. Other eukaryotic cells of interest include plant cells, such as maize, rice, wheat, cotton, soybean, sugarcane, tobacco, and arabidopsis; fish, algae, fungi (Penicillium, Fusarium, Aspergillus, Podospora, Neurospora) , insects, yeasts (Picchia and Saccharomyces).

The choice of host will depend on a number of factors, depending on the intended use of the engineered host, including pathogenicity, substrate range, environmental hardiness, presence of key intermediates, ease of genetic manipulation, and likelihood of promiscuous transfer of genetic information to other organisms. Particularly advantageous hosts are E. coli, lactobacilli, Streptomycetes, Actinomycetes and filamentous fungi.

The breeding procedure starts with at least two substrates, which generally show substantial sequence identity to each other (i.e., at least about 50%, 70%, 80% or 90% sequence identity) but differ from each other at certain positions. The difference can be any type of mutation, for example, substitutions, insertions and deletions. Often, different segments differ from each other in perhaps 5-20 positions. For recombination to generate increased diversity relative to the starting materials, the starting materials must differ from each other in at least two nucleotide positions. That is, if there are only two substrates, there should be at least two divergent positions. If there are three substrates, for example, one substrate can differ from the second as a single position, and the second can differ from the third at a different single position. The starting DNA segments can be natural variants of each other, for example, allelic or species variants. The segments can also be from nonallelic genes showing some degree of structural and is usually functional relatedness (e.g., different genes within a superfamily such as the immunoglobulin superfamily). The starting DNA segments can also be induced variants of each other. For example, one DNA segment can be produced by error- prone PCR replication of the other, or by substitution of a mutagenic cassette. Induced mutants can also be prepared by propagating one (or both) of the segments in a mutagenic strain. In these situations, strictly speaking, the second DNA segment is not a single segment but a large family of related segments. The different segments forming the starting materials are often the same length or substantially the same length. However, this need not be the case; for example; one segment can be a subsequence of another. The segments can be present as part of larger molecules, such as vectors, or can be in isolated form. The starting DNA segments are recombined by any of the recursive sequence recombination formats described above to generate a diverse library of recombinant DNA segments.

Such a library can vary widely in size from having fewer than to more than 10^5, 10^7, or 10^9 members. In general, the starting segments and the recombinant libraries generated include full-length coding sequences and any essential regulatory sequences, such as a promoter and polyadenylation sequence, required for expression. However, if this is not the case, the recombinant DNA segments in the library can be inserted into a common vector providing the missing sequences before performing screening/selection.

If the recursive sequence recombination format employed is an in vivo format, the library of recombinant DNA segments generated already exists in a cell, which is usually the cell type in which expression of the enzyme with altered substrate specificity is desired. If recursive sequence recombination is performed in vitro, the recombinant library is preferably introduced into the desired cell type before screening/selection. The members of the recombinant library can be linked to an episome or virus before introduction or can be introduced directly. In some embodiments of the is invention, the library is amplified in a first host, and is then recovered from that host and introduced to a second host more amenable to expression, selection, or screening, or any other desirable parameter. The manner in which the library is introduced into the cell type depends on the DNA-uptake characteristics of the cell type, e.g., having viral receptors, being capable of conjugation, or being naturally competent. If the cell type is insusceptible to natural and chemical-induced competence, but susceptible to electroporation, one would usually employ electroporation. If the cell type is insusceptible to electroporation as well, one can employ biolistics. The biolistic PDS-1000 Gene Gun (Biorad, Hercules, CA) uses helium pressure to accelerate DNA-coated gold or tungsten microcarriers toward target cells.

The process is applicable to a wide range of tissues, including plants, bacteria, fungi, algae, intact animal tissues, tissue culture cells, and animal embryos. One can employ electronic pulse delivery, which is essentially a mild electroporation format for live tissues in animals and patients. Zhao, Advanced Drug Delivery Reviews 17:257-262 (1995). After introduction of the library of recombinant DNA genes, the cells are optionally propagated to allow expression of genes to occur.

4.6.1.1.2 SELECTION AND SCREENING

Screening is, in general, a two-step process in which one first determines which cells do and do not express a screening marker and then physically separates the cells having the desired property. Selection is a form of screening in which identification and physical separation are achieved simultaneously, for example, by expression of a selectable marker, which, in some genetic circumstances, allows cells expressing the marker to survive while other cells die (or vice versa). Screening markers include, for example, luciferase, ß-galactosidase, and green fluorescent protein.

Screening can also be done by observing such aspects of growth as colony size, halo formation, etc. Additionally, screening for production of a desired compound, such as a therapeutic drug or "designer chemical" can be accomplished by observing binding of cell products to a receptor or ligand, such as on a solid support or on a column. Such screening can additionally be accomplished by binding to antibodies, as in an ELISA. In some instances the screening process is preferably automated so as to allow screening of suitable numbers of colonies or cells. Some examples of automated screening devices include fluorescence activated cell sorting, especially in conjunction with cells immobilized in agarose (see Powell et. al. Bio/Technology 8:333-337 (1990); Weaver et. al. Methods 2:234- 247 (1991)), automated ELISA assays, scintillation proximity assays (Hart, H.E. et al., Molecular Immunol. 16:265-267 (1979)) and the formation of fluorescent, coloured or UV absorbing compounds on agar plates or in microtitre wells (Krawiec, S., Devel. Indust. Microbiology 31:103-114 (1990)).

Selectable markers can include, for example, drug, toxin resistance, or nutrient synthesis genes. Selection is also done by such techniques as growth on a toxic substrate to select for hosts having the ability to detoxify a substrate, growth on a new nutrient source to select for hosts having the ability to utilize that nutrient source, competitive growth in culture based on ability to utilize a nutrient source, etc.

In particular, uncloned but differentially expressed proteins (e.g., those induced in response to new compounds, such as biodegradable pollutants in the medium) can be screened by differential display (Appleyard et al. Mol. Gen. Gent. 247:338-342 (1995)). Hopwood (Phil Trans R. Soc. Lond B 324:549-562) provides a review of screens for antibiotic production. Omura (Microbio. Rev. 50:259-279 (1986) and Nisbet (Ann Rev. Med. Chem. 21:149-157 (1986)) disclose screens for antimicrobial agents, including supersensitive bacteria, detection of beta-lactamase and D,D-carboxypeptidase inhibition, beta-lactamase induction, chromogenic substrates and monoclonal antibody screens.

Antibiotic targets can also be used as screening targets in high throughput screening.

Antifungals are typically screened by inhibition of fungal growth. Pharmacological agents can be identified as enzyme inhibitors using plates containing the enzyme and a chromogenic substrate, or by automated receptor assays. Hydrolytic enzymes (e.g., proteases, amylases) can be screened by including the substrate in an agar plate and scoring for a hydrolytic clear zone or by using a colorimetric indicator (Steele et al. Ann. Rev. Microbiol. 45:89-106 (1991)). This can be coupled with the use of stains to detect the effects of enzyme action (such as congo red to detect the extent of degradation of celluloses and hemicelluloses).

Tagged substrates can also be used. For example, lipases and esterases can be screened using different lengths of fatty acids linked to umbelliferyl. The action of lipases or esterases removes this tag from the fatty acid, resulting in a quenching or enhancement of umbelliferyl fluorescence. These enzymes can be screened in microtiter plates by a robotic device.

4.6.1.1.3 FACS

Fluorescence activated cell sorting (FACS) methods are also a powerful tool for selection/screening. In some instances a fluorescent molecule is made within a cell (e.g., green fluorescent protein). The cells producing the protein can simply be sorted by FACS. Gel microdrop technology allows screening of cells encapsulated in agarose microdrops (Weaver et al. Methods 2:234-247 (1991)). In this technique products secreted by the cell (such as antibodies or antigens) are immobilized with the cell that generated them. Sorting and collection of the drops containing the desired product thus also collects the cells that made the product, and provides a ready source for the cloning of the genes encoding the desired functions. Desired products can be detected by incubating the encapsulated cells with fluorescent antibodies (Powell et al. Bio/Technology 8:333-337 (1990)). FACS sorting can also be used by this technique to assay resistance to toxic compounds and antibiotics by selecting droplets that contain multiple cells (i.e., the product of continued division in the presence of a cytotoxic compound; Goguen et al. Nature 363:189-190 (1995)). This method can select for any enzyme that can change the fluorescence of a substrate that can be immobilized in the agarose droplet.

4.6.1.1.4 REPORTER MOLECULE

In some embodiments of the invention, screening can be accomplished by assaying reactivity with a reporter molecule reactive with a desired feature of, for example, a gene product. Thus, specific functionalities such as antigenic domains can be screened with antibodies specific for those determinants.

4.6.1.1.5 CELL-CELL INDICTAOR

In other embodiments of the invention, screening is preferably done with a cell-cell indicator assay. In this assay format, separate library cells (Cell A, the cell being assayed) and reporter cells (Cell B, the assay cell) are used.

Only one component of the system, the library cells, is allowed to evolve. The screening is generally carried out in a two-dimensional immobilized format, such as on plates. The products of the metabolic pathways encoded by these genes (in this case, usually secondary metabolites such as antibiotics, polyketides, carotenoids, etc.) diffuse out of the library cell to the reporter cell. The product of the library cell may affect the reporter cell in one of a number of ways.

The assay system (indicator cell) can have a simple readout (e.g., green fluorescent protein, luciferase, ß- galactosidase) which is induced by the library cell product but which does not affect the library cell. In these examples the desired product can be detected by colorimetric changes in the reporter cells adjacent to the library cell.

4.6.1.1.6 FEEDBACK MECHANISM

In other embodiments, indicator cells can in turn produce something that modifies the growth rate of the library cells via a feedback mechanism. Growth rate feedback can detect and accumulate very small differences. For example, if the library and reporter cells are competing for nutrients, library cells producing compounds to inhibit the growth of the reporter cells will have more available nutrients, and thus will have more opportunity for growth. This is a useful screen for antibiotics or a library of polyketide synthesis gene clusters where each of the library cells is expressing and exporting a different polyketide gene product.

4.6.1.1.7 SECRETION

Another variation of this theme is that the reporter cell for an antibiotic selection can itself secrete a toxin or antibiotic that inhibits growth of the library cell. Production by the library cell of an antibiotic that is able to suppress growth of the reporter cell will thus allow uninhibited growth of the library cell.

Conversely, if the library is being screened for production of a compound that stimulates the growth of the reporter cell (for example, in improving chemical syntheses, the library cell may supply nutrients such as amino acids to an auxotrophic reporter, or growth factors to a growth-factor- dependent reporter. The reporter cell in turn should produce a compound that stimulates the growth of the library cell. Interleukins, growth factors, and nutrients are possibilities. Further possibilities include competition based on ability to kill surrounding cells, positive feedback loops in which the desired product made by the evolved cell stimulates the indicator cell to produce a positive growth factor for cell A, thus indirectly selecting for increased product formation.

In some embodiments of the invention it can be advantageous to use a different organism (or genetic background) for screening than the one that will be used in the final product. For example, markers can be added to DNA constructs used for recursive sequence recombination to make the microorganism dependent on the constructs during the improvement process, even though those markers may be undesirable in the final recombinant microorganism.

Likewise, in some embodiments it is advantageous to use a different substrate for screening an evolved enzyme than the one that will be used in the final product. For example, Evnin et al. (Proc. Natl. Acad. Sci. U.S.A. 87:6659-6663 (1990)) selected trypsin variants with altered substrate specificity by requiring that variant trypsin generate an essential amino acid for an arginine auxotroph by cleaving arginine - naphthylamide. This is thus a selection for arginine-specific trypsin, with the growth rate of the host being proportional to that of the enzyme activity.

The pool of cells surviving screening and/or selection is enriched for recombinant genes conferring the desired phenotype (e.g. altered substrate specificity, altered biosynthetic ability, etc.). Further enrichment can be obtained, if desired, by performing a second round of screening and/or selection without generating additional diversity.

The recombinant gene or pool of such genes surviving one round of screening/selection forms one or more of the substrates for a second round of recombination. Again, recombination can be performed in vivo or in vitro by any of the recursive sequence recombination formats described above.

If recursive sequence recombination is performed in vitro, the recombinant gene or genes to form the substrate for recombination should be extracted from the cells in which screening/selection was performed. Optionally, a subsequence of such gene or genes can be excised for more targeted subsequent recombination. If the recombinant gene(s) are contained within episomes, their isolation presents no difficulties. If the recombinant genes are chromosomally integrated, they can be isolated by amplification primed from known sequences flanking the regions in which recombination has occurred. Alternatively, whole genomic DNA can be isolated, optionally amplified, and used as the substrate for recombination. Small samples of genomic DNA can be amplified by whole genome amplification with degenerate primers (Barrett et al. Nucleic Acids Research 23:3488-3492 (1995)). These primers result in a large amount of random 3' ends, which can undergo homologous recombination when reintroduced into cells.

If the second round of recombination is to be performed in vivo, as is often the case, it can be performed in the cell surviving screening/selection, or the recombinant genes can be transferred to another cell type (e.g., a cell is type having a high frequency of mutation and/or recombination). In this situation, recombination can be effected by introducing additional DNA segment(s) into cells bearing the recombinant genes. In other methods, the cells can be induced to exchange genetic information with each other by, for example, electroporation. In some methods, the second round of recombination is performed by dividing a pool of cells surviving screening/selection in the first round into two subpopulations. DNA from one subpopulation is isolated and transfected into the other population, where the recombinant gene(s) from the two subpopulations recombine to form a further library of recombinant genes. In these methods, it is not necessary to isolate particular genes from the first subpopulation or to take steps to avoid random shearing of DNA during extraction. Rather, the whole genome of DNA sheared or otherwise cleaved into manageable sized fragments is transfected into the second subpopulation. This approach is particularly useful when several genes are being evolved simultaneously and/or the location and identity of such genes within chromosome are not known.

The second round of recombination is sometimes performed exclusively among the recombinant molecules surviving selection. However, in other embodiments, additional substrates can be introduced. The additional substrates can be of the same form as the substrates used in the first round of recombination, i.e., additional natural or induced mutants of the gene or cluster of genes, forming the substrates for the first round. Alternatively, the additional substrate(s) in the second round of recombination can be exactly the same as the substrate(s) in the first round of replication.

After the second round of recombination, recombinant genes conferring the desired phenotype are again selected. The selection process proceeds essentially as before. If a suicide vector bearing a selective marker was used in the first round of selection, the same vector can be used again. Again, a cell or pool of cells surviving selection is selected. If a pool of cells, the cells can be subject to further enrichment.

4.6.1.2 GENERAL METHODS

4.6.1.2.1 IN VITRO

In Vitro Formats one format for recursive sequence recombination in vitro is illustrated herein. The initial substrates for recombination are a pool of related sequences. The X's show where the sequences diverge. The sequences can be DNA or RNA and can be of various lengths depending on the size of the gene or DNA fragment to be recombined or stochastic &/or non-stochastic mutagenized. Preferably the sequences are from 50 bp to 100 kb.

The pool of related substrates are converted into overlapping fragments, e.g., from about 5 bp to 5 kb or more, as shown herein. Often, the size of the fragments is from about 10 bp to 1000 bp, and sometimes the size of the DNA fragments is from about 100 bp to 500 bp. The conversion can be effected by a number of different methods, such as DNaseI or RNase digestion, random shearing or partial restriction enzyme digestion.

Alternatively, the conversion of substrates to fragments can be effected by incomplete PCR amplification of substrates or PCR primed from a single primer. Alternatively, appropriate single-stranded fragments can be generated on a nucleic acid synthesizer. The concentration of nucleic acid fragments of a particular length and sequence is often less than 0. 1 % or 1% by weight of the total nucleic acid. The number of different specific nucleic acid fragments in the mixture is usually at least about 100, 500 or 1000.

The mixed population of nucleic acid fragments are converted to at least partially single-stranded form. Conversion can be effected by heating to about 80C to 100C, more preferably from 90C to 96 C, to form single-stranded nucleic acid fragments and then reannealing. Conversion can also be effected by treatment with single-stranded DNA binding protein or recA protein. Single-stranded nucleic acid fragments having regions of sequence identity with other single-stranded nucleic acid fragments can then be reannealed by cooling to 4C to 75C, and preferably from 40C to 65C. Renaturation can be accelerated by the addition of polyethylene glycol (PEG), other volume-excluding reagents or salt. The salt concentration is preferably from 0 mM to 200 mM more preferably the salt concentration is from 10 mM to 100 mM. The salt may be KC1 or NaCl. The concentration of PEG is preferably from 0% to 20%, more preferably from 5% to 10%. The fragments that reanneal can be from different substrates as shown herein. The annealed nucleic acid fragments are incubated in the presence of a nucleic acid polymerase, such as Taq or Klenow, or proofreading polymerases, such as pfu or pwo, and dNTP's (i.e. dATP, dCTP, dGTP and dTTP). If regions of sequence identity are large, Taq polymerase can be used with an annealing temperature of between 45-65C. If the areas of identity are small, Klenow polymerase can be used with an annealing temperature of between 20-30T (Stemmer, Proc. Natl. Acad. Sci. USA (1994), supra). The polymerase can be added to the random nucleic acid fragments prior to annealing, simultaneously with annealing or after annealing.

The process of denaturation, renaturation and incubation in the presence of polymerase of overlapping fragments to generate a collection of polynucleotides containing different permutations of fragments is sometimes referred to as stochastic &/or non-stochastic mutagenesis of the nucleic acid in vitro. This cycle is repeated for a desired number of times. Preferably the cycle is repeated from 2 to 100 times, more preferably the sequence is repeated from 10 to 40 times. The resulting nucleic acids are a family of double-stranded polynucleotides of from about 50 bp to about 100 kb, preferably from 500 bp to 50 kb, as shown herein. The population represents variants of the starting substrates showing substantial sequence identity thereto but also diverging at several positions. The population has many more members than the starting substrates. The population of fragments resulting from stochastic &/or non-stochastic mutagenesis is used to transform host cells, optionally after cloning into a vector.

4.6.1.2.1.1 FULL LENGTH SEQUENCES

In a variation of in vitro stochastic &/or non-stochastic mutagenesis, subsequences of recombination substrates can be generated by amplifying the full-length sequences under conditions which produce a substantial fraction, typically at least 20 percent or more, of incompletely extended amplification products. The amplification products, including the incompletely extended amplification products are denatured and subjected to at least one additional cycle of reannealing and amplification. This variation, wherein at least one cycle of reannealing and amplification provides a substantial fraction of incompletely extended products, is termed "stuttering." In the subsequent amplification round, the incompletely extended products anneal to and prime extension on different sequence-related template species.

4.6.1.2.1.2 OVERLAPPING SINGLE STRANDED DNA FRAGMENTS

In a further variation, at least one cycle of amplification can be conducted using a collection of overlapping single-stranded DNA fragments of related sequence, and different lengths. Each fragment can hybridize to and prime polynucleotide chain extension of a second fragment from the collection, thus forming sequence-recombined polynucleotides. In a further variation, single-stranded DNA fragments of variable length can be generated from a single primer by Vent DNA polymerase on a first DNA template. The single stranded DNA fragments are used as primers for a second, Kunkel-type template, consisting of a uracil-containing circular single-stranded DNA. This results in multiple substitutions of the first template into the second (see Levichkin et al. Mol. Biology 29:572-577 (1995)).

4.6.1.2.1.3 GENE CLUSTERS

Gene clusters such as those involved in polyketide synthesis (or indeed any multi-enzyme pathways catalyzing is analogous metabolic reactions) can be recombined by recursive sequence recombination even if they lack DNA sequence homology. Homology can be introduced using synthetic oligonucleotides as PCR primers. In addition to the specific sequences for the gene being amplified, all of the primers used to amplify one type of enzyme (for example the acyl carrier protein in polyketide synthesis) are synthesized to contain an additional sequence of 20-40 bases 51 to the gene (sequence A) and a different 20-40 base sequence 31 to the gene (sequence B). The adjacent gene (in this case the keto- synthase) is amplified using a 51 primer which contains the complementary strand of sequence B (sequence B'), and a 31 primer containing a different 20-40 base sequence (C). Similarly, primers for the next adjacent gene (keto- reductases) contain sequences C' (complementary to C) and D. If 5 different polyketide gene clusters are being stochastic &/or non-stochastic mutagenized, all five acyl carrier proteins are flanked by sequences A and B following their PCR amplification. In this way, small regions of homology are introduced, making the gene clusters into site specific recombination cassettes. Subsequent to the initial amplification of individual genes, the amplified genes can then be mixed and subjected to primerless PCR. Sequence B at the 3' end of all of the five acyl carrier protein genes can anneal with and prime DNA synthesis from sequence BI at the 5' end of all five keto reductase genes. In this way all possible combinations of genes within the cluster can be obtained. Oligonucleotides allow such recombinants to be obtained in the absence of sufficient sequence homology for recursive sequence recombination described above. Only homology of function is required to produce functional gene clusters.

4.6.1.2.1.4 MULTI SUBUNIT ENZYMES

This method is also useful for exploring permutations of any other multi-subunit enzymes. An example of such enzymes composed of multiple polypeptides that have shown novel functions when the subunits are combined in novel ways are dioxygenases. Directed recombination between the four protein subunits of biphenyl and toluene dioxygenases produced functional dioxygenases with increased activity against trichloroethylene (Furukawa et. al. J. Bacteriol. 176: 2121-2123 (1994)). This combination of subunits from the two dioxygenases could also have been produced by cassette- stochastic &/or non-stochastic mutagenesis of the dioxygenases as described above, followed by selection for degradation of trichloroethylene.

In some polyketide synthases, the separate functions of the acyl carrier protein, keto-synthase, keto- reductase, etc. reside in a single polypeptide. In these cases domains within the single polypeptide may be stochastic &/or non-stochastic mutagenized, even if sufficient homology does not exist naturally, by introducing regions of homology as described above for entire genes. In this case, it may not be possible to introduce additional flanking sequences to the domains, due to the constraint of maintaining a continuous open reading frame.

Instead, groups of oligonucleotides are synthesized that are homologous to the 3' end of the first domain encoded by one of the genes to be stochastic &/or non-stochastic mutagenized, and the 5' ends of the second domains encoded by all of the other genes to be stochastic &/or non-stochastic mutagenized together. This is repeated with all domains, thus providing sequences that allow recombination between protein domains while maintaining their order.

4.6.1.2.1.5 CASSETTE-BASED

The cassette-based recombination method can be combined with recursive sequence recombination by including gene fragments (generated by DNase, physical shearing, DNA stuttering, etc.) for one or more of the genes. Thus, in addition to different combinations of entire genes within a cluster (e.g., for polyketide synthesis), individual genes can be stochastic &/or non-stochastic mutagenized at the same time (e.g., all acyl carrier protein genes can also be provided as fragmented DNA), allowing a more thorough search of sequence space.

4.6.1.2.1.6 IN VITRO WHOLE GENOME STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS

The stochastic &/or non-stochastic mutagenesis of large DNA sequences, such as eukaryotic chromosomes, is difficult by prior art in vitro stochastic &/or non-stochastic mutagenesis methods. A method for overcoming this limitation is described herein.

The cells of related eukaryotic species are gently lysed and the intact chromosomes are liberated. The liberated chromosomes are then sorted by FACS or similar method (such as pulse field electrophoresis) with chromosomes of similar size being sequestered together. Each size fraction of the sorted chromosomes generally will represent a pool of analogous chromosomes, for example the Y chromosome of related mammals. The goal is to isolate intact chromosomes that have not been irreversibly damaged.

The fragmentation and stochastic &/or non-stochastic mutagenesis of such large complex pieces of DNA employing DNA polymerases is difficult and would likely introduce an unacceptably high level of random mutations. An alternative approach that employs restriction enzymes and DNA ligase provides a feasible less destructive solution. A chromosomal fraction is digested with one or more restriction enzymes that recognize long DNA sequences (about 15 - 20bp), such as the intron and intein encoded endonucleases (I-Ppo 1, I-Ceu I, PI-Psp 1, PI- Tli 1, PI-Sce I (VDE).

These enzymes each cut, at most, a few times within each chromosome, resulting in a combinatorial mixture of large fragments, each having overhanging single stranded termini that are complementary to other sites cleaved by the same enzyme.

The digest is further modified by very short incubation with a single stranded exonuclease. The polarity of the nuclease chosen is dependent on the single stranded overhang resulting from the restriction enzyme chosen. 5'-3' exonuclease for 3'-overhangs, and 3'-5'- exonuclease for 5'overhangs. This digestion results in significantly long regions of ssDNA overhang on each dsDNA termini. The purpose of this incubation is to generate regions of DNA that define specific regions of DNA where recombination can occur. The fragments are then incubated under condition where the ends of the fragments anneal with other fragments having homologous ssDNA termini. Often, the two fragments annealing will have originated from different chromosomes and in the presence of DNA ligase are covalently linked to form a chimeric chromosome. This generates genetic diversity mimicking the crossing over of homologous chromosomes. The complete ligation reaction will contain a combinatorial mixture of all possible ligations of fragments having homologous overhanging termini. A subset of this population will be complete chimeric chromosomes.

To screen the stochastic &/or non-stochastic mutagenized library, the chromosomes are delivered to a suitable host in a manner allowing for the uptake and expression of entire chromosomes. For example, YACs (yeast artificial chromosomes) can be delivered to eukaryotic cells by protoplast fusion.

Thus, the reassemble library could be encapsulated in liposomes and fused with protoplasts of the appropriate host cell. The resulting transformants would be propagated and screened for the desired cellular improvements. Once an improved population was identified, the chromosomes would be isolated, stochastic &/or non-stochastic mutagenized, and screened recursively.

4.6.1.2.1.7 WHOLE GENOME STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS OF NATURALLY COMPETENT MICROORGANISMS

Natural competence is a phenomenon observed for some microbial species whereby individual cells take up DNA from the environment and incorporate it into their genome by homologous recombination. Bacillus subtilis and Acetinetobacter Spp. are known to be particularly efficient at this process. A method for the whole genome stochastic &/or non-stochastic mutagenesis of these and analogous organisms is described employing this process.

One goal of whole genome stochastic &/or non-stochastic mutagenesis is the rapid accumulation of useful mutations from a population of individual strains into one superior strain. If the organisms to be evolved are naturally competent, then a split pooled strategy for the recursive transformation of naturally competent cells with DNA originating from the pool will effect this process. An example procedure is as follows.

A population of naturally competent organisms that demonstrates a variety of useful traits (such as increased protein secretion) is identified. The strains are pooled, and the pool is split. One half of the pool is used as a source of gDNA, while the other is used to generate a pool of naturally competent cells.

The competent cells are grown in the presence of the pooled gDNA to allow DNA uptake and recombination. Cells of one genotype uptake and incorporate gDNA from cells of a different type generating cells having chimeric genomes. The result is a population of cells representing a combinatorial mixture of the genetic variations originating in the original pool. These cells are pooled again and transformed with the same source of DNA again. This process is carried out recursively to increase the diversity of the genomes of cells resulting from transformation. Once sufficient diversity has been generated, the cell population is screened for new chimeric organisms demonstrating desired improvements.

This process is enhanced by increasing the natural competence of the host organism. COMS is a protein that, when expressed in B. subtilis, enhances the efficiency of natural competence mediated transformation more than an order of magnitude.

It was demonstrated that approximately 100% of the cells harboring the plasmid pCOMS uptake and recombine genomic DNA fragments into their genomes. In general, approximately 10% of the genome is recombined into any given transformed cell. This observation was demonstrated by the following.

A strain of B. subtilis pCOMS auxotrophic for two nutritional markers was transformed with genomic DNA (gDNA) isolated from a prototrophic strain of the same organism. 10% of the cells exposed to the DNA were prototrophic for one of the two nutrient markers. The average size of the DNA strand taken up by B. subtilis is approximately 50kb or about 2% of the genome. Thus 1 of every ten cells had recombined a marker that was represented 1 in every fifty molecules of uptaken gDNA. Thus, most of the cells take up and recombine with approximately five 50kb molecules or 10% of the genome. This method represents a powerful tool for rapidly and efficiently recombining whole microbial genomes.

In the absence of pCOMS, only 0.3% of the cells prepared for natural competency uptake and integrate a specific marker. This suggested that about 15% of the cells actually underwent recombination with a single genomic fragment. Thus, a recursive transformation strategy as described above produces a whole genome stochastic &/or non-stochastic mutagenized library, even in the absence of pCOMS. In the absence of pCOMS, however, the complex genomes will represent a smaller, but still screenable percentage of the transformed or stochastic &/or non-stochastic mutagenized population.

4.6.1.2.1.8 CONGRESSION

Congression is the integration of two independent unlinked markers into a cell. 0.3% of naturally competent B. subtilis cells integrate a single marker (described above). Of these, about 10% have taken up an additional marker. Thus, if one selects or screens for the integration of one specific marker, 10% of the resulting population will have integrated another specific marker. This provides a way of enriching for specific integration events.

For example, if one is looking for the integration of a gene for which there is no easy screen or selection, it will exist as 0.3% of the cell population. If the population is first selected for a specific integration event, then the desired integration will be found in 10% of the population. This represents a significant (about 30-fold) enrichment for the desired event. This enrichment is defined as the "congression effect." The congression effect is not influenced by the presence of pCOMS, thus the "pCOMS effect" is simply to increase the percentage of naturally competent cells that are truly naturally competent from about 15% in its absence to 100% in its presence. All competent cells still uptake about the same amount of DNA or about 10% of the Bacillus genome.

The congression effect can be used in the following examples to enhance whole genome stochastic &/or non-stochastic mutagenesis as well, as the targeted integration of stochastic &/or non-stochastic mutagenized genes to the chromosome.

4.6.1.2.1.9 B.SUBTILIS STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS

A population of B. subtilis cells having desired properties are identified, pooled and stochastic &/or non-stochastic mutagenized as described above with one exception: once the pooled population is split, half of the population is transformed with an antibiotic selection marker that is flanked by sequence that targets its integration and disruption of a specific nutritional gene, for example, one involved in amino biosynthesis. Transformants resistant to the drug are auxotrophic for that nutrient. The resistant population is pooled and grown under conditions rendering them naturally competent (or optionally first transformed with pCOMS).

The competent cells are then transformed with gDNA isolated from the original pool, and prototrophs are selected. The prototrophic population will have undergone recombination with genomic fragments encoding a functional copy of the nutritional marker, and thus will be enriched for cells having undergone recombination at other genetic loci by the congression effect.

4.6.1.2.1.10 TARGETING OF GENES AND GENE LIBRARIES TO THE CHROMOSOME

It is useful to be able to efficiently deliver genes or gene libraries directly to a specific location in a cells chromosome. As above, target cells are transformed with a positive selection marker flanked by sequences that target its homologous recombination into the chromosome. Selected cells harboring the marker are made naturally competent (with or without pCOMS, but preferably the former) and transformed with a mixture of two sets of DNA fragments. The first set contains a gene or a stochastic &/or non-stochastic mutagenized library of genes each flanked with sequence to target its integration to a specific chromosomal loci. The second set contains a positive selection marker (different from that first integrated into the cells) flanked by sequence that will target its integration and replacement of the first positive selection marker.

Under optimal conditions, the mixture is such that the gene or gene library is in molar excess over the positive selection marker. Transformants are then selected for cells containing the new positive marker. These cells are enriched for cells having integrated a copy of the desired gene or gene library by the congression effect and can be directly screened for cells harboring the gene or gene variants of interest. This process was carried out using PCR fragments <10kb, and it was found that, employing the congression effect, a population can be enriched such that 50% of the cells are congregants. Thus, one in two cells contained a gene or gene variant.

Alternatively, the expression host can be absent of the first positive selection marker, and the competent cells are transformed with a mixture of the target genes and a limiting amount of the first positive selection marker fragment. Cells selected for the positive marker are screened for the desired properties in the targeted genes. The improved genes are amplified by the PCR, stochastic &/or non-stochastic mutagenized again, and then returned to the original host again with the first positive selection marker. This process is carried out recursively until the desired function of the genes are obtained. This process obviates the need to construct a primary host strain and the need for two positive markers.

4.6.1.2.1.11 CONJUGATION-MEDIATED GENETIC EXCHANGE

Conjugation can be employed in the evolution of cell genomes in several ways. Conjugative transfer of DNA occurs during contact between cells. See Guiney(1993)in: Bacterial Conjugation (Clewell, ed., Plenum Press, New York), pp. 75-104; Reimmann & Haas in Bacterial Conjugation (Clewell, ed., Plenum Press, New York 1993), at pp. 137-188 (incorporated by reference in their entirety for all purposes). Conjugation occurs between many types of gram negative bacteria, and some types of gram positive bacteria. Conjugative transfer is also known between bacteria and plant cells (Agrobacterium tumefaciens) or yeast. As discussed in patent 5,837,458, the genes responsible for conjugative transfer can themselves be evolved to expand the range of cell types (e.g., from bacteria to mammals) between which such transfer can occur.

Conjugative transfer is effected by an origin of transfer (oriT) and flanking genes (MOB A, B and C, and 15-25 genes, termed tra, encoding the structures and enzymes necessary for conjugation to occur. The transfer origin is defined as the site required in cis for DNA transfer. Tra genes include tra A, B, C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, T, U,V,W, X,Y,Z,virAB(allelesl-ll),C,D,E,G,IHF,andFinOP. Tra genes can be expressed in cis or trans to oriT. Other cellular enzymes, including those of the RecBCD pathway, RecA, SSB protein, DNA gyrase, DNA poll, and DNA ligase, are also involved in conjugative transfer. RecE or recF pathways can substitute for RecBCD.

One structural protein encoded by a tra gene is the sex pilus, a filament constructed of an aggregate of a single polypeptide protruding from the cell surface. The sex pilus binds to a polysaccharide on recipient cells and forms a conjugative bridge through which DNA can transfer. This process activates a site-specific nuclease encoded by a MOB gene, which specifically cleaves DNA to be transferred at oriT. The cleaved DNA is then threaded through the conjugation bridge by the action of other tra enzymes.

Mobilizable vectors can exist in episomal form or integrated into the chromosome. Episomal mobilizable vectors can be used to exchange fragments inserted into the vectors between cells. Integrated mobilizable vectors can be used to mobilize adjacent genes from the chromosome.

4.6.1.2.1.12 USE OF INTEGRATED MOBILIZED VECTORS TO PROMOTE EXCHANGE OF GENOMIC DNA

The F plasmid of E. coli integrates into the chromosome at high frequency and mobilizes genes unidirectional from the site of integration (Clewell, 1993, supra; Firth et al., in Escherichia coli and Salmonella Cellular and Molecular Biology 2, 23 77-2401 (1996); Frost et al., Microbiol. Rev. 58, 162-210 (1994)). Other mobilizable vectors do not spontaneously integrate into a host chromosome at high efficiency, but can be induced to do so by growth under particular conditions (e.g., treatment with a mutagenic agent, growth at a nonpermissive temperature for plasmid replication). See Reimann & Haas in Bacterial Conjugation (ed. Clewell, Plenum Press, NY 1993), Ch. 6. Of particular interest is the IncP group of conjugal plasmids which are typified by their broad host range (Clewell, 1993, supra. Donor "male" bacteria which bear a chromosomal insertion of a conjugal plasmid, such as the E. coli F factor can efficiently donate chromosomal DNA to recipient "female" enteric bacteria which lack F (F-). Conjugal transfer from donor to recipient is initiated at oriT. Transfer of the nicked single strand to the recipient occurs in a 5' to 3' direction by a rolling circle mechanisms which allows mobilization of tandem chromosomal copies. Upon entering the recipient, the donor strand is discontinuously replicated. The linear, single-stranded donor DNA strand is a potent substrate for initiation of recA-mediated homologous recombination within the recipient. Recombination between the donor strand and recipient chromosomes can result in the inheritance of donor traits. Accordingly, strains which bear a chromosomal copy of F are designated Hfr (for high frequency of recombination) (Low, 1996 in Escherichia coli and Salmonella Cellular and Molecular Biology Vol. 2, pp. 2402- 2405; Sanderson, in Escherichia coli and Salmonella Cellular and Molecular Biology 2, 2406-2412 (1996)).

The ability of strains with integrated mobilizable vector to transfer chromosomal DNA provides a rapid and efficient means of exchanging genetic material between a population of bacteria thereby allowing combination of positive mutations and dilution of negative mutations. Such stochastic &/or non-stochastic mutagenesis methods typically start with a population of strains with an integrated mobilizable vector encompassing at least some genetic diversity.

The genetic diversity can be the result of natural variation, exposure to a mutagenic agent or introduction of a fragment library. The population of cells is cultured without selection to allow genetic exchange, recombination and expression of recombinant genes. The cells are then screened or selected for evolution toward a desired property. The population surviving selection/screening can then be subject to a further round of stochastic &/or non-stochastic mutagenesis by IVR-mediated genetic exchange, or otherwise.

The natural efficiency of Hfr and other strains with integrated mob vectors as recipients of conjugal transfer can be improved by several means. The relatively low recipient efficiency of natural BFR strains is attributable to the products of traS and traT genes of F (Clewell, 1993, supra; Firth et al., 1996, supra.- Frost et al., 1994, supra; Achtman et al., J Mol. Biol. 138, 779-795 (1980). These products are localized to the inner and outer membranes of F+ strains, respectively, where they serve to inhibit redundant matings between two strains which are both capable of donating DNA. The effects of traS and traT, and cognate genes in other conjugal plasmids, can be eliminated by use of knockout cells incapable of expressing these enzymes or reduced by propagating cells on a carbon- limited source. (Peters et al., J Bacteriol., 178, 3037-3043 (1996)).

In some methods, the starting population of cells has a mobilizable vector integrated at different genomic sites. Directional transfer from oriT typically results in more frequent inheritance of traits proximal to oriT. This is because mating pairs are fragile and tend to dissociate (particularly when in liquid medium) resulting in the interruption of transfer.

In a population of cells having a mobilizable vector integrated at different sites, chromosomal exchange occurs in a more random fashion. Kits of Hfr strains are available from the E coli. Genetic Stock Center and the Salmonella Genetic Stock Centre (Frost et al., 1994, supra).

Alternatively, a library of strains with oriT at random sites and orientations can be produced by insertion mutagenesis using a transposon which bears oriT. The use of a transposon bearing an oriT [e.g., the Tn5-oriT described by Yakobson EA, et al. J. Bacteriol. 1984 Oct; 160(l): 451-453] provides a quick method of generating such a library. Transfer functions for mobilization from the transposon-borne oriT sites are provided by a helper vector in trans. It is possible to generate similar genetic constructs using other sequences known to one of skill as well.

In one aspect, a recursive scheme for genomic stochastic &/or non-stochastic mutagenesis using Tn-oriT elements is provided. A prototrophic bacterial strain or set of related strains bearing a conjugal plasmid, such as the F fertility factor or a member of the IncP group of broad host range plasmids is mutagenized and screened for the desired properties. Individuals with the desired properties are mutagenized with a Tn-oriT element and screened for acquisition of an auxotrophy (e.g., by replica-plating to a minimal and complete media) resulting from insertion of the Tn-oriT element in any one of many biosynthetic gene scattered across the genome. The resulting auxotrophs are pooled and allowed to mate under conditions promoting male-to-male matings, e.g., during growth in close proximity on a filter membrane. Note that transfer functions are provided by the helper conjugal plasmid present in the original strain set. Recombinant transconjugants are selected on minimal medium and screened for further improvement.

Optionally, strains bearing integrated mobilizable vectors are defective in mismatch repair gene(s). Inheritance of donor traits which arise from sequence heterologies increases in strains lacking the methyl-directed mismatch repair system. Optionally, the gene products which decrease recombination efficiency can be inhibited by small molecules.

Intergenic conjugal transfer between species such as E. coli and Salmonella typhimurium, which are 20% divergent at the DNA level, is also possible if the recipient strain is mutH, mutL or mutS (see Rayssiguier et al., Nature 342, 396-401 (1989)). Such transfer can be used to obtain recombination at several points as shown by the following example.

One example uses an S. typhimurium Hfr donor strain having markers thr557 at map position 0, pyrF2690 at 33 min, serA13 at 62 min and hfrK5 at 43 min. MutS +/-, F- E. coli. recipient strains had markers pyrD68 at 21 min aroC355 at 51 min, ilv3164 at

85 min and mutS215 at 59 min. The triauxotrophic S. typhimurium Hfr donor and isogenic mutS+/- triauxotrophic E. coli recipient were inoculated into 3 ml of Lb broth and shaken at 37C until fully grown. 100 ul of the donor and each recipient were mixed in 10 ml fresh LB broth, and then deposited to a sterile Millipore 0.45 uM HA filter using a Nalgene 250 ml reusable filtration device. The donor and recipients alone were similarly diluted and deposited to check for reversion. The filters with cells were placed cell-side-up on the surface of an LB agar plate which was incubated overnight at 37C. The filters were removed with the aid of a sterile forceps and placed in a sterile 50 ml tube containing 5 ml of minimal salts broth. Vigorous vortexing was used to wash the cells from the filters. 100 ul of mating mixtures, as well as donor and recipient controls were spread to LB for viable cell counts and minimal glucose supplemented with either two of the three recipient requirements for single recombinant counts, one of the three requirements for double recombinant counts, or none of the three requirements for triple recombinant counts. The plates were incubated for 48 hr at 37C after which colonies were counted.

Frequencies are further enhanced by increasing the ratio of donor to recipient cells, or by repeatedly mating the original donor strains with the previously generated recombinant progeny.

4.6.1.2.1.13 INTRODUCTION OF FRAGMENTS BY CONJUGATION

Sobilizable vectors can also be used to transfer fragment libraries into cells to be evolved. This approach is particularly useful in situations in which the cells to be evolved cannot be efficiently transformed directly with the fragment library but can undergo conjugation with primary cells that can be transformed with the fragment library. DNA fragments to be introduced into host cells encompasses diversity relative to the host cell genome. The diversity can be the result of natural diversity or mutagenesis.

The DNA fragment library is cloned into a mobilizable vector having an origin of transfer. Some such vectors also contain mob genes although alternatively these functions can also be provided in trans. The vector should be capable of efficient conjugal transfer between primary cells and the intended host cells. The vector should also confer a selectable phenotype. This 96 phenotype can be the same as the phenotype being evolved or can be conferred by a marker, such as a drug resistance marker. The vector should preferably allow self-elimination in the intended host cells thereby allowing selection for cells in which a cloned fragment has undergone genetic exchange with a homologous host segment rather than duplication. Such can be achieved by use of vector lacking an origin of replication functional in the intended host type or inclusion of a negative selection marker in the vector.

One suitable vector is the broad host range conjugation plasmid described by Simon et al., Bio/Technology 1, 784-791 (1983); TrieuCuot et al., Gene 102, 99-104 (1991); Bierman et al., Gene 116, 43-49 (1992). These plasmids can be transformed into E. coli and then force-mated into bacteria that are difficult or impossible to transform by chemical or electrical induction of competence. These plasmids contain the origin of the IncP plasmid, oriT Mobilization functions are supplied in trans by chromosomally- integrated copies of the necessary genes. Conjugal transfer of DNA can in some cases be assisted by treatment of the recipient (if gram-positive) with sub-inhibitory concentrations of penicillins (Trieu-Cuot et al., 1993 FEMS Microbiol. Lett. 109, 19-23). To increase diversity in populations, recursive conjugal mating prior to screening is performed.

Cells that have undergone allelic exchange with library fragments can be screened or selected for evolution toward a desired phenotype. Subsequent rounds of recombination can be performed by repeating the conjugal transfer step. The library of fragments can be fresh or can be obtained from some (but not all) of the cells surviving a previous round of selection/screening. Conjugation-mediated stochastic &/or non-stochastic mutagenesis can be combined with other methods of stochastic

&/or non-stochastic mutagenesis.

4.6.1.2.1.14 GENETIC EXCHANGE PROMOTED BY TRANSDUCING PHAGE IN CELLS SUSEPTIBLE TO PHAGE

Phage transduction can include the transfer, from one cell to another, of nonviral genetic material within a viral coat (Masters, in Escherichia coli and Salmonella Cellular and Molecular Biology 2, 2421-2442 (1996). Perhaps the two best examples of generalized transducing phage are bacteriophages P I and P22 of E. coli and S. typhimurium, respectively. Generalized transducing bacteriophage particles are formed at a low frequency during lytic infection when viral-genome-sized, doubled-stranded fragments of host (which serves as donor) chromosomal DNA are packaged into phage heads. Promiscuous high transducing (HT) mutants of bacteriophage P22 which efficiently package DNA with little sequence specificity have been isolated. Infection of a susceptible host results in a lysate in which up to 50% of the phage are transducing particles. Adsorption of the generalized transducing particle to a susceptible recipient cell results in the injection of the donor chromosomal fragment. RecA-mediated homologous recombination following injection of the donor fragment can result in the inheritance of donor traits. Another type of phage which achieves quasi random insertion of DNA into the host chromosome is Mu. For an overview of Mu biology, see, Groisman (1991) in Methods in Enzymology v. 204. Mu can generate a variety of chromosomal rearrangements including deletions, inversions, duplications and transpositions. In addition, elements which combine the features of P22 and Mu are available, including Mud-P22, which contains the ends of the Mu genome in place of the P22 att site and int gene. See, Berg, supra.

Generalized transducing phage can be used to exchange genetic material between a population of cells encompassing genetic diversity and susceptible to infection by the phage. Genetic diversity can be the result of natural variation between cells, induced mutation of cells or the introduction of fragment libraries into cells. DNA is then exchanged between cells by generalized transduction. If the phage does not cause lysis of cells, the entire population of cells can be propagated in the presence of phage. If the phage results in lytic infection, transduction is performed on a split pool basis. That is, the starting population of cells is divided into two. One subpopulation is used to prepare transducing phage. The transducing phage are then infected into the other subpopulation. Preferably, infection is performed at high multiplicity of phage per cell so that few cells remain uninfected. Cells surviving infection are propagated and screened or selected for evolution toward a desired property. The pool of cells surviving screening/selection can then be stochastic &/or non-stochastic mutagenized by a further round of generalized transduction or by other stochastic &/or non-stochastic mutagenesis methods. Recursive split pool tranduction is optionally performed prior to selection to increase the diversity of any population to be screened.

The efficiency of the above methods can be increased by reducing infection of cells by infectious (nontransducing phage) and by reducing lysogen formation. The former can be achieved by inclusion of chelators of divalent cations, such as citrate and EDTA in culture media. Tail defective transducing phages can be used to allow only a single round of infection.

Divalent cations are required for phage absorption and the inclusion of chelating agents therefore provides a means of preventing unwanted infection. Integration defective (int) derivatives of generalized transducing phage can be used to prevent lysogen formation. In a further variation, host cells with defects in mismatch repair gene(s) can be used to increase recombination between transduced DNA and genomic DNA.

4.6.1.2.1.15 USE OF LOCKED IN PROPHAGES TO FACILITATE DNA STOCHASTIC &/OR NON-STOCHASTIC MUTAGENESIS

The use of a hybrid, mobile genetic element (locked-in prophages) as a means to facilitate whole genome stochastic &/or non-stochastic mutagenesis of organisms using phage transduction as a means to transfer DNA from donor to recipient is a preferred embodiment. One such element (Mud-P22) based on the temperate Salmonella phage P22 has been described for use in genetic and physical mapping of mutations. See, Youderian et al. (1988) Genetics 118:581 - 592, and Benson and Goldman (1992) J Bacteriol. 174(5):1673-1681. Individual Mud-P22 insertions package specific regions of the Salmonella chromosome into phage P22 particles.

Libraries of random Mud-P22 insertions can be readily isolated and induced to create pools of phage particles packaging random chromosomal DNA fragments. These phage particles can be used to infect new cells and transfer the DNA from the host into the recipient in the process of transduction. Alternatively, the packaged chromosomal DNA can be isolated and manipulated further by techniques such as DNA stochastic &/or non-stochastic mutagenesis or any other mutagenesis technique prior to being reintroduced into cells (especially recD cells for linear DNA) by transformation or electroporation, where they integrate into the chromosome. Either the intact transducing phage particles or isolated DNA can be subjected to a variety of mutagens prior to reintroduction into cells to enhance the mutation rate.

Mutator cell fines such as mutD can also be used for phage growth. Either method can be used recursively in a process to create genes or strains with desired properties. E. coli cells carrying a cosmid clone of Salmonella LPS genes are infectable by P22 phage. It is possible to develop similar genetic elements using other combinations of transposable elements and bacteriophages or viruses as well. P22 is a lambdoid phage that packages its DNA into pstochastic &/or non-stochastic mutagenized phage particles (heads) by a "headful" mechanism. Packaging of phage DNA is initiated at a specific site (pac) and proceeds unidirectionally along a linear, double stranded normally concatameric molecule. When the phage head is full (about 43 kb), the DNA strand is cleaved, and packaging of the next phage head is initiated. Locked-in or excision-defective P22 prophages, however, initiate packaging at their pac site, and then proceed unidirectionally along the chromosome, packaging successive headfuls of chromosomal DNA (rather than phage DNA). When these transducing phages infect new Salmonella cells they inject the chromosomal DNA from the original host into the recipient cell, where it can recombine into the chromosome by homologous recombination creating a chimeric chromosome. Upon infection of recipient cells at a high multiplicity of infection, recombination can also occur between incoming transducing fragments prior to recombination into the chromosome.

Integration of such locked-in P22 prophages at various sites in the chromosome allows flanking regions to be amplified and packaged into phage particles. The Mud-P22 mobile genetic element contains an excision-defective P22 prophage flanked by the ends of phage/transposon Mu. The entire Mud-P22 element can transpose to virtually any location in the chromosome or other episome (eg. F', BAC clone) when the Mu A and B proteins are provided in trans.

A number of embodiments for this type of genetic element are available. In one example, the locked in prophage are used as generalized transducing phage to transfer random fragments of a donor chromosome into a recipient. The Mud-P22 element acts as a transposon when Mu A and B transposase proteins are provided in trans and integrate copies of itself at random locations in the chromosome. In this way, a library of random chromosomal Mud-P22 insertions can be generated in a suitable host. When the Mud-P22 prophages in this library are induced, random fragments of chromosomal DNA will be packaged into phage particles. When these phages infect recipient cells, the chromosomal DNA is injected and can recombine into the chromosome of the recipient. These recipient cells are screened for a desired property and cells showing improvement are then propagated.

The process can be repeated, since the Mud-P22 genetic element is not transferred to the recipient in this process. Infection at a high multiplicity allows for multiple chromosomal fragments to be injected and recombined into the recipient chromosome. Locked in prophages can also be used as specialized transducing phage.

Individual insertions near a gene of interest can be isolated from a random insertion library by a variety of methods. Induction of these specific prophages results in packaging of flanking chromosomal DNA including the gene(s) of interest into phage particles. Infection of recipient cells with these phages and recombination of the packaged DNA into the chromosome creates chimeric genes that can be screened for desired properties. Infection at a high multiplicity of infection can allow recombination between incoming transducing fragments prior to recombination into the chromosome.

These specialized transducing phage can also be used to isolate large quantities of high quality DNA containing specific genes of interest without any prior knowledge of the DNA sequence. Cloning of specific genes is not required. Insertion of such an element nearby a biosynthetic operon for example allows for large amounts of DNA from that operon to be isolated for use in DNA stochastic &/or non-stochastic mutagenesis (in vitro and/or in vivo), cloning, sequencing, or other uses as set forth herein. DNA isolated from similar insertions in other organisms containing homologous operons are optionally mixed for use in family stochastic &/or non-stochastic mutagenesis formats as described, herein, in which homologous genes from different organisms (or different chromosomal locations within a single species, or both). Alternatively, the transduced population is recursively transduced with pooled transducing phage or new transducing phage generated from the previously transduced cells. This can be carried out recursively to optimize the diversity of the genes prior to stochastic &/or non-stochastic mutagenesis.

Phage isolated from insertions in a variety of strains or organisms containing homologous operons are optionally mixed and used to coinfect cells at a high MOI allowing for recombination between incoming transducing fragments prior to